Antero Karvonen

FORMS OF DETERMINATION IN NATURAL AND ARTIFICIAL SYSTEMS

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2018

TIIVISTELMÄ

Karvonen, Antero
Forms of determination in natural and artificial systems
Jyväskylä: Jyväskylän yliopisto, 2018, 93 s.
Kognitiotiede, pro gradu -tutkielma)
Ohjaaja: Saariluoma, Pertti

Tutkielman tarkoitus on arvioida autonomisen teknologian ja keinoälyn taustalle vallitsevia olettamuksia perusteanalyyttisestä näkökulmasta käyttäen autonomisia laivoja kontekstina. Teoreettinen tutkielma on kriittinen, mutta motiiveiltaan rakentava: mikäli olettamuksien taustalla vallitsevia konseptuaalisia ongelmia kyetään tunnistamaan, niitä voidaan myös pyrkiä paikkaamaan. Tämä tavoite on oletettavasti pitkän tähtäimen kehitystä ja tavoitteena sellaisenaan tutkielman ulkopuolella. Siitä huolimatta teknologian perustavaa laatua olevien rajoitteiden tunnistaminen on ensiarvoisen tärkeää aikana, jolloin kehitys on nopeaa ja riskit todellisia. Tämä mahdollistaa ihmisen ja autonomisen teknologian vuorovaikutukseen liittyvien kysymyksien asemoimisen rationaaliselle pohjalle ja auttaa tunnistamaan molempien vahvoja ja heikkoja puolia. Tiivistettynä tutkielman tavoite on kuvata niitä vaatimuksia, joita autonominen teknologia asettaa pyrkiessään korvaamaan inhimillistä tiedonkäsittelyä teknisessä järjestelmässä. Tämä pyritään kytkemään kriittiseen keskusteluun tietokoneiden ja laskennallisten toimenpiteiden perustavaa laatua olevien ominaisuuksien kyvystä saavuttaa näitä vaatimuksia. Kriittisestä keskustelusta seuraa kaksi keskeistä johtopäätöstä. 1. Inhimillisen tiedonkäsittelyn korvaaminen kokonaisuutena on nykyisten teknisten järjestelmien mahdollisuuksien ulkopuolella. Tekniset järjestelmät pystyvät korvaamaan ja tukemaan inhimillistä tiedonkäsittelyä hyvin määritellyissä ja spesifeissä tehtävissä, kuten nytkin, mutta toistaiseksi erilaisen tiedon integroiminen merkityksellisiksi kokonaisuuksiksi ja sen pohjalta kumpuava tavoitteellinen toiminta ei vaikuta teknisten järjestelmien toimintaperiaatteiden valossa realistiselta tavoitteelta. 2. Näin ollen miehittämättömien ja etäohjauksessa toimivien laivajärjestelmien kehittämisessä tulisi kiinnittää erityisesti huomiota siihen, miten etäohjauskeskuksen operaattorit saavat käyttöönsä kaiken sen (tiedostamattoman ja tiedostetun) tiedon jonka pohjalta he tekevät päätöksiä perinteisten laivojen kannella, sikäli kun inhimillinen tiedonkäsittely tulee säilymään välttämättömänä osana laivan toimintaa, ainakin kriittisissä ja haastavissa olosuhteissa.

Asiasanat: autonomiset laivat, autonominen teknologia, keinoäly, perusteanalyysi

ABSTRACT

Karvonen, Antero
Forms of determination in natural and artificial systems
Jyväskylä: University of Jyväskylä, 2018, 93 p.
Cognitive Science, Master's Thesis
Supervisor: Saariluoma, Pertti

This thesis is a theoretical review of a few of the central issues pertaining to autonomous technical artefacts, using ships as a context, and by extension artificial intelligence and cognitive science. The approach is critical on the one hand, and constructive on the other, in that it proceeds from the idea that the central struggles facing AI are not merely technical but conceptual. It is thus an analysis of some of the presuppositions that underlie AI and an attempt at turning attention towards questions that need to be addressed if proper autonomy and intelligence are to be achieved in artefacts. If the conceptual problems identified are true, it means we may also attempt to address them. This fix as such is beyond the scope of this thesis, but at a time of rapid change and real risks, understanding the foundational limitations of technology is of paramount importance. This should also serve to position questions relating to human-technology interaction on a rational basis, and help to identify strengths and weaknesses of both. In brief, the goal of this thesis is to outline the requirements posed by autonomous technology insofar as it seeks to, or must, replace human information-processing from a technical system. We seek to connect this into a critical and foundational discussion on the limitations of computers and computations in fulfilling those requirements. Two main conclusions flow from the critical discussion. First, in settings that include dynamic and unpredictable characteristics, the replacement human information-processing as whole is beyond the capacities of current technical solutions. They can, as they are now, be used to support and even replace some facets of human cognition in specific and well-defined tasks. But so far the integration of different information into meaningful wholes, that goal-directed and context-sensitive action requires, are seen as an unrealistic goal for technical artefacts in light of the operating principles of computers. In our view, this is not yet a mere technical problem, but a conceptual and analytical one. Second, given that humans will remain a necessary component for quasi-autonomous ship operations in the near future, extreme care should be put into the design of the unmanned operations and specifically the remote operation centers, such that the necessary information (tacit or explicit) by which decisions are made on the ship's bridge will translate into the remote operation center.

Keywords: autonomous ships, autonomous technology, artificial intelligence, foundational analysis

FIGURES

TABLES

TABLE OF CONTENTS

1 PREFACE

Ours is a time of tremendous excitement over intelligent technology, as anyone following the news, academia, or industry is likely to admit. Artificial intelligence, robotics, or machine learning of some sort or another are seen as the next big thing. But for anyone knowledgeable on the history of AI, excitement is nothing new, and neither are correspondingly large disappointments. Whether or not this time we have the right concrete and conceptual tools to make truly intelligent technology a reality is, at the moment, an open question. But history as our guide, there is certain cause for consideration. This thesis represents one attempt at mapping some parts of a complex and controversial landscape by way of a foundational analysis of the presuppositions, concepts, and principles from which AI in its' various forms emerges, and to extend the discourse towards broader topics of multiple realizability and forms of determination, and back to engineering design. It is thus intended as a general theoretical review of the requirements of autonomy vis-à-vis human information-processing and the corresponding challenges this poses for technical systems. This means that much detail is left out, as well as some regions of the landscape. If some of the critiques hold, however, it means that truly intelligent machines remain more than a mere technical challenge. It means that there are analytical and conceptual problems which need to be addressed. These questions in turn begin approaching the very core and foundations of not only AI, but cognitive science and beyond.

2 INTRODUCTION

Spearheaded in the public imagination by self-driving cars, and riding on a more general wave of excitement over intelligent technology, there is a proliferation of attempts to create autonomous technology. The latest in this wave are autonomous maritime vessels (Levander, 2017; Rolls-Royce, 2016). The technical, legal, social, security, and business -related problems and requirements the quest for autonomous ships present are significant and intertwined. As an example, whether the technical solutions to problems presented by the task environment of the ship are secure and reliable enough has direct ramifications for the legal and business dimensions of the whole enterprise. An unreliable autonomous ship may not be legally allowed to sail or a high-profile failure may become a PR disaster sinking the business possibilities of such ships, to name a few examples. Such multi-dimensional problems must be tackled from correspondingly varied domains of human understanding. Recently, Luciano Floridi (2017) articulated this in the language of levels of abstraction. It simply means that experts from different domains pick up different patterns of information from the same set of observables[1] - a legal scholar sees very different types of problems than a software engineer, a cognitive scientist yet others. Given that those of us in the enterprise are involved in a quintessentially future-oriented exploration of *possibilities*, rather than establishing or contributing *only* facts about an existing domain, we could say that what different experts most robustly contribute are not (only) answers but *questions*. Questions that both guide and shape the design process and products. Indeed, by identifying the question-solution structure of design activities, Saariluoma, Cañas, and Leikas (2016) suggested a *question-structure* - oriented approach to design. This approach, Life-Based Design, is predicated on the idea that many different technical solutions can satisfy a functional requirement. Thus, an LBD ontology for design is characterized by sets of relevant questions that point to bodies of knowledge and discourse, rather than as a repository of formal facts and relationships.

This thesis is organized around the basic question whether the concepts and principles of artificial intelligence and computer science suffice for proper autonomy. We seek to examine the foundational questions around autonomy, and to critically examine the presuppositions of AI against those requirements. Our attempt is thus critical on the one hand, and constructive on the other. Within the limits of this thesis, we seek to explore the limitations of current conceptualizations in AI by way of a foundational examination of the presuppositions underlying it (Saariluoma, 1997). This is a different approach to a typical approach in AI, namely pure performance in some test or another (for a classic test see Turing, 1950).

Of course, this is not to argue *against* operational tests of autonomy or intelligence. Pure performance is a practical method for measurement, and certainly

[1] Given that autonomous ships do not exist yet, we might more accurately say that they co-create the observables, rather than simply observe them.

necessary from the perspective of actual systems to be deployed. However, given that machines that could generalize beyond whatever domain they were programmed to perform in and not suffer dramatic breakdowns in performance under modified situations have not been achieved (Shoham, Perrault, Brynjolfsson, & Clark, 2017), and even expert opinion is divided on the issue of general machine intelligence (Müller & Bostrom, 2016), the assessment of the foundational issues remains fruitful. Arguably even if an AI system were to be developed that matched all human benchmarks in performance, the foundational issues would remain open for investigation. But here, given that the central practical issues, such as common-sense reasoning (Lake, Ullman, Tenenbaum, & Gershman, 2017; Marcus, 2018; McCarthy, 2007), remain unsolved, we have justification beyond the "merely philosophical" in evaluating the foundational issues.

If we can connect the notion of autonomy with mental determination, intelligence, and cognition, and cast serious doubt on the ability of the fundamental principles of computers in achieving those, then we are in a position to re-evaluate the first principles from which AI should proceed. This has relevance for the philosophy of cognitive science as well. It should be noted, that we are not going to make the case that *human beings* could not possess the ingenuity to cast down into a causal mechanism certain forms of human information-processing. Human ingenuity is not the target of this thesis, but the limits of the fundamental working principles of computers in intrinsically embodying something *like* human ingenuity.

What questions from various levels of abstraction (in the sense of Floridi 2017) and different bodies of human understanding contribute are ways in which the problem can be defined more precisely or more widely. Clarifying and extending the problem of autonomy from the perspective of cognitive science and the rich discourse around artificial intelligence (among other topics) is one of the main goals for this thesis. What follows from such a clarification, should be a tentative mapping of the distance or mismatch between the goal of autonomy and the prevailing state (of technical systems) (Leppänen, 2005). Our results should serve the design processes around autonomous ships from two perspectives. First, identifying limitations of artefacts in achieving human-level performance in certain tasks should inform how the sharing of duties between man and machine should be laid out in the near future. It seems likely that both have their virtues and failings. Second, if the former is done from a foundational, "deep" perspective, it may provide ideas and contribute to the design of intelligent technology itself. Such results are of course highly tentative, and subject not only to the veracity of the findings but also their applicability for practical concerns – there is no doubt that practical life and engineering design will proceed in its' own way regardless. But it should be borne in mind, that if the fundamental principles are ill-conceived for the task at hand then there is no hope for achieving ultimate success.

Essentially the question before us is to examine the fundamental presuppositions, concepts, and argumentation that surround artificial intelligence. What is important is to place the questions marks deep enough, so that the questions and

(possible) answers approach the heart of the matter – only in this way can we avoid getting lost in the foggy details and technical arguments that will inevitably crop up. Thus, we shall not be, other than by way of example, asking whether deep neural nets or more traditional methods of AI are the most suitable for artificial intelligence, nor about the extent to which the systems should learn by example or have pre-built innate mechanisms (Marcus, 2018; Pearl, 2018). We want to ask is the computer and computations a suitable platform for mentality – and thus autonomy proper - to emerge, or is there something to life itself (Rosen, 1999) and the forms of determination it embodies (Bunge, 1979) that are a necessary precondition for proper autonomy, mentality, and intelligence?

2.1 Context and Rationale

This work proceeds within and is partly funded by the Tekes-funded DIMECC Design for value program that seeks to "enable the best possible use of digital disruption in supply chains" (DIMECC, 2017). One central aspect of the digital supply chains is envisaged to be unmanned autonomous ships. Some reports place such ships in 2025 (Levander, 2017), a mere seven years away at the time of writing. In the D4value program, research is being conducted from multiple points of view simultaneously. The task of Jyväskylä University in the program is to study human-work interaction in this changing context.

The rationale for the thesis is to provide a foundational viewpoint for the discourse. This extends the thesis beyond autonomous ships and towards autonomous technology in general, and in the way we shall proceed, towards the very core questions of cognitive science and artificial intelligence. Examination of these questions is crucial for the development of technology and understanding its' limitations, perhaps even transcending them. Furthermore, by understanding, or at least raising the crucial questions surrounding the limitations of technical artefacts in approximating human thinking feeds directly into the way in which future human-work interaction in an ecosystem of quasi-autonomous artefacts and human operators should be structured.

2.2 Method

Issuing from the author's interests on the one hand, and the seeming lack of clarity in the foundations of what autonomy entails on the other (Boden, 2008), this thesis is of rather broad scope, made even more so by the inclusion of artificial intelligence (which is itself foundationally unclear, see Saariluoma & Rauterberg, 2015; 2016) as a necessary corollary for autonomy. As a kind of excavation of presuppositions, the research process itself was a recursive circling around the four main topics reflected by the four main sections: autonomy; artificial intelligence;

multiple realizability; and engineering design. We can't hope to exhaustively examine or describe each of the domains, but offer a particular set of ideas gleaned from each of the domains, and hope that the results have coherence and illustrate some of the difficulties inherent in the quest for truly autonomous technology.

3 AUTONOMY IN SHIPS

The purpose of this section is to orient ourselves around the topic of autonomy in the context of maritime vessels, and in general. We will attempt to define some key terms, to illustrate some of the ways autonomy has been measured, identify some its' key dimensions, and explore the relationships between system, autonomy, and environment. The goal of the section is to somewhat clarify the problem of autonomy and establish a link between human autonomy and mental capacities, and correspondingly, artefact autonomy and artificial intelligence.

3.1 General Introduction

This work proceeds within the general framework of the Tekes-funded DIMECC Design for value program that seeks to enable the best possible use of digital disruption in supply chains (DIMECC, 2017). One central aspect of the digital supply chains is envisaged to be unmanned autonomous ships (Levander, 2017; Rolls-Royce, 2016), and it is around this topic we will proceed. In the Design for value program, research is being conducted from multiple points of view simultaneously. A legal scholar sees an autonomous ship from a different perspective than an engineer or a business person. The observations of this thesis will emerge from an interdisciplinary framework which seeks to advance understanding of mind and intelligence in humans and in general: cognitive science (Frankish & Ramsey, 2012; Thagard, 2005). The task of Jyväskylä University in the program is to study human-work interaction. This thesis contributes to this question through the following logic:

1. Correctly formulated general requirements for autonomy define the research space. This is identical to the current situation because the man-ship system is autonomous in the sense relevant here;
2. Given the goal of autonomy, the limitations of technological solutions in solving for those requirements define how that space is carved up or shared between humans and artefacts;
3. This defines the roles of man and machine, and ultimately allows for research into human-work interaction to proceed in a principled manner. This part is obviously subject to change as technology progresses.
4. However, understanding the foundational limitations of (current) technical artefacts, if any, may point the way towards transcending them and thus contribute to the design ontology of autonomous technology.

The precursor for the Design4Value program was the AAWA (Advanced Autonomous Waterborne Applications) project (Rolls-Royce, 2016). This will be

our starting point. The vision for autonomous ships that emerged from the project was a system of systems in which autonomy is gradually approached by rolling out autonomous functionalities in ships while simultaneously having the possibility of placing the ships under remote control from a shore control center when the situation demands it. The vision is relatively modest and realistic which means the assumption is not that the ship has the capacity to deal with even particularly many or complex situations. Indeed, such moments may be relatively few in the timespan of a navigation from port-to-port if significant parts of the journey take place over the open ocean. To concretize the vision a bit, let us paraphrase a typical journey for an unmanned containership under remote control as the AAWA report saw it (Rolls-Royce, 2016, p. 8-12).

3.1.1 An Example Journey

In this example journey, the ship is unmanned with the ability to perform certain tasks such as follow a navigational path and avoid simple collisions. It is tethered via data link to a remote operator who can monitor the ship and assist and take control as needed. The ship is a containership and will travel from one port to another via an open sea route.

3.1.1.1 Voyage Planning and Initiation

From the AAWA perspective, this aspect will be mostly under the control and discretion of the operator. The operator will plan the journey, perhaps with technological assistance, and set beforehand which legs of the journey will be in autonomous mode (such as long stretches of open ocean) and which will be under remote control (such as departure from a congested port). The operator will set navigational strategies for each leg. They will include fallback strategies which determine how the ship is to proceed if, for example, it encounters a situation which is unexpected and contact the operator, if failing that, proceeding to a next waypoint, failing that holding position, or navigating to previous waypoint. This sequence of fallback strategies would only be implemented if connectivity breaks down and the ship encounters an unexpected situation.

3.1.1.2 Unmooring and Manouvering out of the Harbor

Departure from a harbor will likely require human operators for the time being, especially if the harbor is congested and contains a mixture of different vessel types and enough dynamic and complex characteristics to render safe autonomous operations difficult. This will depend strongly on the systems available at the harbor, for example if they have a mooring and unmooring systems that can be automated and are suitable for the particular ship in question. The control of the ship may be either quasi-direct, meaning the joystick and throttle commands from the remote center will directly control the actuators in the ship, or once-removed so that and the ship is only given manual waypoints, taking care of movements autonomously.

3.1.1.3 Operating in the Open Sea

Once the harbor area is cleared and the shipping lanes are relatively traffic-free, the journey begins according to the plan set by the operator prior to departure. In this mode, the ship makes its way from waypoint to waypoint, while monitoring its' local environment through its "sensory" systems such as radar, lidar, or cameras. The systems interpreting the local environment should alert the remote operator if situations demanding attention arise. In normal operations, as the journey is proceeding according to the plan, the ship is quite autonomous. Thus, the autonomy-level is adjusted according to situation, it is not a feature which applies to the ship as a constant property. Interaction with the remote operator is of different types. If the ship if executing a maneuver between two waypoints in a manner which does not exceed some pre-specified margins, it may only notify the operator which then can veto if the situation demands it. Of course, sometimes situations are likely to arise where the path planning and collision avoidance modules are unable to unambiguously solve a certain situation, or may contradict each other in terms of goals. Such complex situations demand much from the ship and by extension the programmers, because it requires dynamic conflict-resolution between different goals and weighted criteria. Below is a capture from the AAWA report, illustrating different ship-operator interaction scenarios.

Figure 1 Illustration of different navigational scenarios from the perspective of the remote operator (Rolls-Royce, 2016)

3.1.1.4 Port Approach and Docking

The last part is largely a reverse of the departure. It should be noted that pilotage systems and the profession will require modification with the arrival of

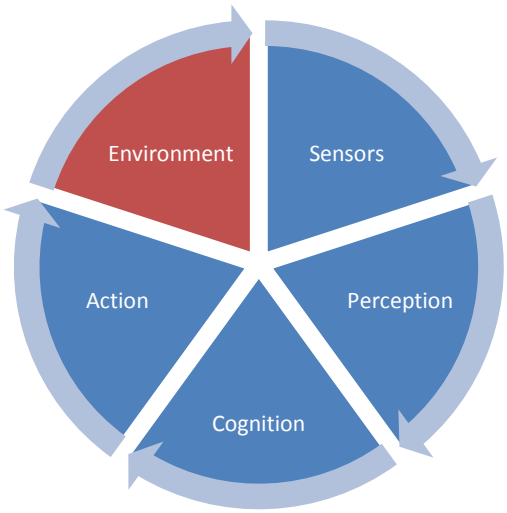
autonomous ships. In general, much work is required from the perspective of harbors as well including infrastructure and personnel training.

3.1.2 General Architecture for Autonomous Artefacts

When it comes to architectures for autonomous vehicles, a 2011 review of the state of the art in decision-making identified challenges for any systematic review (Veres, Molnar, Lincoln, & Morice, 2011). For example, there are almost as many architectures as there are vehicles, and hundreds of sub-problems in navigation and control of all kinds. See also Siegwart, Nourbakhsh, & Scaramuzza (2011). The following is a general breakdown of a typical architecture for a mobile autonomous artefact, based on Siegwart et al. (2011), Veres et al. (2011), Schiaretti, Chen, & Negenborn (2017a, 2017b), which is capitulated in spirit in the AAWA (Advanced Autonomous Waterborne Applications) whitepaper (Rolls-Royce, 2016) and its' Autonomous Navigation System (ANS).

The four fundamental layers of sensory, perceptual, motor, and cognitive capacities form a necessary unity for the achievement of autonomy in an artefact (Siegwart et al., 2011). Without sensors the artefact is blind; without perception it can't make sense of its environment; without motor capacities it is immobile; and without cognition it can't make decisions or solve problems. These functional necessities are to a degree commensurate with the functions identified by Parasuraman, Sheridan, & Wickens (2000): information acquisition; information analysis; decision and action selection; and action implementation - or sensory, perceptual, cognitive, and motor in our analysis. The figure below summarizes the key aspects.

Figure 2 General Architecture for an Autonomous Artefact



3.1.2.1 Techniques and Implementation Methods

The task of navigating from one point to another can be broken down to local and global path planning problems and methods. A review by Polvara, Sharma, Wan, Manning, & Sutton (2017) concluded that while global path planning is hardly a technical problem assuming accurate cartographic information, the same is not true for local problems: i.e. obstacle detection and avoidance. On the one hand, obstacle detection depends on robust and reliable sensor-systems. On the other, avoidance of obstacles, especially if there are many and they are moving, requires complex decision-making and evaluation, while taking into account vessel dynamics, weather conditions, and the maritime rules of the road (COLREGS) (Polvara et al., 2017). Perhaps most concerning of all, many solutions have only been tested in computer simulations, which, while appropriate for testing and development, will always lack the actual complexity potential in maritime situations.

3.1.3 Autonomous Navigation System

The architecture of the autonomous navigation system (ANS) as planned in the AAWA project (Rolls Royce, 2016, p.20) follows the general architecture outlined before to a large degree. See figure below.

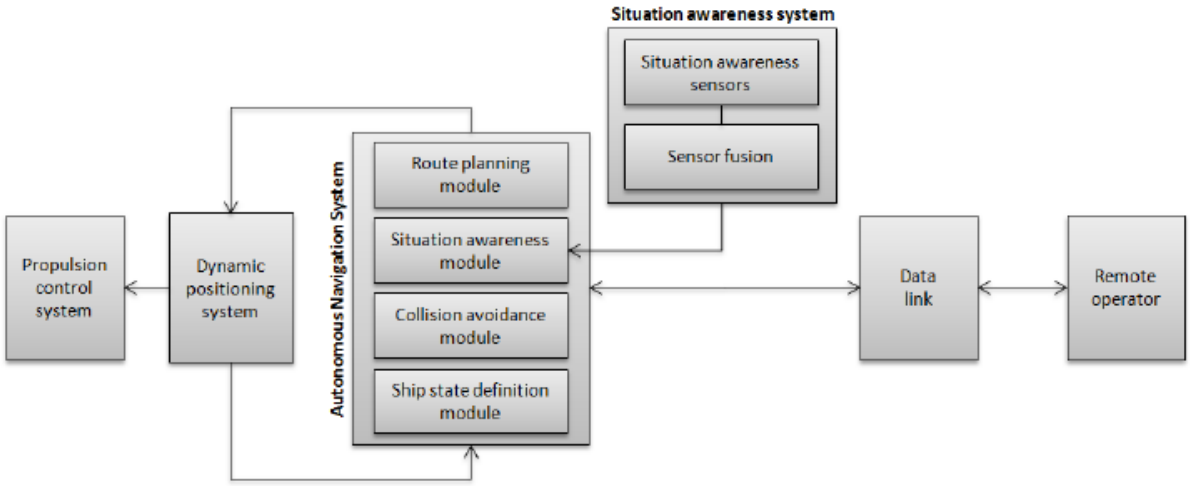


Figure 3 Architecture of the Autonomous Navigation System (ANS) (Rolls-Royce, 2016)

The basic elements that are required for all autonomous mobile artefacts are there: the dynamic positioning system and the propulsion control system account for motor capacities; situation awareness sensors and sensor fusion account for sensory and perception layers; and the autonomous navigation system with route planning, situation awareness, collision avoidance, and ship state definition modules accounts (to some extent) for cognition. The devil is of course in the details and whether the functional breakdown matches actual systems is an open

question, and of course this schematic is merely an illustrative plan. This will our reference point as to the perceptual and cognitive systems of the autonomous ship. Let us now examine the key modules in more detail.

3.1.3.1 Situation Awareness

The purpose of the situation awareness (SA) module is to monitor and extract relevant information from the environment via sensory systems for use by the collision avoidance (CA) module. Its' job is to supply a *local* map of the immediate environment and show obstacles around the ship. The sensors by which it makes sense of the environment may include LIDAR (short range radar), which can provide accurate range and velocity information. Cameras work better for classification of objects. The key, for the AAWA project, has been sensor fusion, which essentially means the transformation of data obtained from various sensors into a representation usable both for the ship's internal systems, but also for the remote human operator. Situation awareness thus goes two ways, into the ship and into the operation center. The SA module is the eyes of the ship, metaphorically speaking.

3.1.3.2 Collision Avoidance

Collision Avoidance (CA) is tasked with dealing with situations in the *local* environment, and is thus intimate with the SA module described before. It takes the view from the bridge of the ship, metaphorically speaking. In the ANS architecture, it ties in with the DP (dynamic positioning) module (which is already in use on ships) which acts as the last link between the ANS and the actuators of the ship. The CA module is perhaps misleadingly called that, because it's role is to also navigate the ship under normal conditions, following the route delivered by the path planning module which we will discuss next. As said, the CA module assesses the risks presented by obstacles and objects in the environment and executes manouvers within suitable parameters (while consulting with the ship state definition module which is the integrator of all ship information).

3.1.3.3 Route Planning

Route planning (RP) takes the bird's eye view of the whole journey. Even if the operator has set the waypoints manually, the planned path is located in the RP module. The route follows shipping lanes when available, and avoids known obstacles based on electronic chart data. It consists of waypoints, headings, and speeds for the ship. The route is the strategic level of the journey, containing as mentioned before, multiple fallback strategies for particular legs of the journey. The tactical manouvers are made by the CA module and the remote operator.

3.1.3.4 Ship State Definition Module or "Virtual Captain"

The highest level in the module hierarchy is the ship state definition (SSD) module or "virtual captain". This module gathers information from all other modules, and informs the remote operator of the state of the ship. It also controls which mode the ship is in, semi-autonomous, full-autonomous, or under remote control. It is thus a meta-awareness module for the entire ship.

3.1.4 Possible Future Development Pattern

As the previous illustrations have made clear, the human "recedes", as it were, from the ship (see figure below). More generally speaking, the word ship is here co-extensive with the technical artefacts that make up the immediate ship, and includes all such aspects of the ship such as radars, sonars, LIDARs, and in a sense most importantly, the computer systems that are seeking to replace human thought and intentionality from the equation: essentially the ANS and its' modules in the description before.

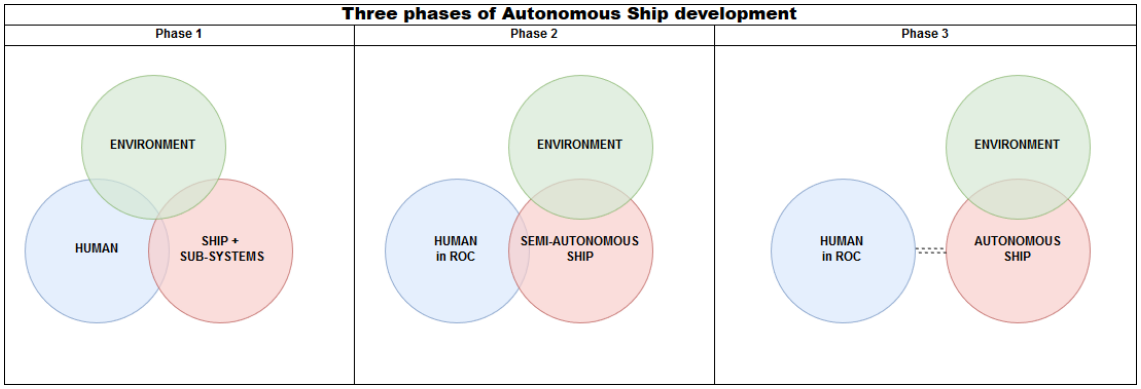


Figure 4 Three phases of autonomous ship development

The figure above seeks to illustrate one way of conceptualizing the possible future development of autonomous ships.

We start with *phase one*, the current situation. Here, the humans on the ship, the ship and its subsystems, and the environment form a spatially intimate whole and all three domains have immediate interaction among each other.

The second phase entails the removal of man from the immediate ship and into a Remote Operation Center (ROC) or Shore Control Center (SCC). This presupposes that the ship has both adequate mastery over many typical maritime situations, or at least the capacity to stop and wait for human intervention, and that the necessary certainty has been achieved in the connectivity between the ROC and the ship, and that the ROC has been properly designed and manned.

Finally, in *phase three* the ship has mastered the capacity to deal with almost all maritime situations, and the remote operation center is used to take over only in very rare circumstances, and is used rather to supervise and manage a larger

fleet of autonomous ships. Our focus in this thesis will be to look ahead into this phase.

From this we can see that there are two central technological lines of development. The first is the building up of capacities in the ship to autonomously handle situations. The second is the development of principles, practices, usability, and controllability of the ships from the remote operation center. We will proceed from the idea that the space circumscribed by the capacities of the current man-ship system is populated by requirements from the perspective of design. How, and how well the requirements are met currently set the standards that a future autonomous ship must equal or surpass. What the requirements should do is act as goals to guide design and innovation. The idea is that all shifts within the space of requirements for autonomous ships are the transfers of functions and capacities from the human to the ship's domain, probably through intermediate stages – a gradient of autonomy if you will. The types of requirements we will focus on are of the kind marked by intelligence, and mostly identifiable with the bridge of the ship. This is not to say that the maintenance of the ship engines, electronic systems, or the responsibilities of the steward do not require intelligence, indeed they may be much harder tasks to automate than others. But the chief aim of autonomous ships, as the discussion before illustrated, is to replace humans in the tasks most pertinent to the navigation, steering, and corresponding decision-making, and those are largely identifiable with the bridge of the ship, and specifically with the first, second, and third mates, and ultimately the captain.

In the AAWA project, a realistic and relatively modest starting point has been assumed, which seems smart given that no fully robust solution exists for autonomous navigation, especially in the local aspect of it (Polvara et al., 2017). In this thesis, we will "look ahead" into future scenarios for two reasons. First, we are interested in limitations of technology, given that it feeds directly into the primary research question of Jyväskylä University in the program, human-work interaction. Second, of more significance to the field of cognitive science and AI in general is a discussion on such foundational limitations. Thus, we seek to contribute both to the research questions for our institution, as well as to the field at large, and of course to the future development of autonomous shipping and autonomous technology in general.

3.1.5 The Automatic Steering of Ships

The automatic steering of ships is hardly a new phenomenon. It's root trace back at least to the seminal work by Nicolas Minorsky (1922) in the early 20th century. By observing the behavior of experienced helmsmen in maintaining a heading, Minorsky (1922) was able to capture its' essential features in a mathematical formulation, which today forms the basis of a common class of industrial control, the PID controller (Bennett, 1984). A PID controller works by a feedback loop where a setpoint, a desired value, is compared against the actual performance value of the system in question. The latter value is the sum of proportional,

integrative, and derivative terms (hence PID), which correspond to the error between the setpoint and the desired point, the integration of past error values over time, and the derivation of a best guess for the future based on current rate of change.

What makes Minorsky's work interesting from our perspective is, on the one hand, the method of observation of actual helmsmen by which he arrived at his conclusion, and on the other, his keen awareness of nonlinearities in systems (a ship with six degrees of freedom influenced by complex environmental factors), which allowed him to abstract out the variables that were general enough, and relevant for the problem he was attempting to solve (Bennett, 1984). By abstracting out the nonlinear forces that act on a ship that form the cause of the disturbance in most cases, and anchoring his solution to two known variables, a setpoint and an actual point, he was able to form a causally closed loop between two known variables that as a result is a general solution to the problem of maintaining a course in a relatively universal way. What is further captured by his formula is history and anticipation. Essentially Minorsky's (1922; Bennett, 1984) solution is a part of what might now be called control theory, and its' precursor, cybernetics (Wiener, 1985; Checkland, 1994). Indeed, the term cybernetics comes from the Greek word meaning steersman, and taking as its' subject matter all possible machines or systems that require control of some sort, seeks to abstract out the general principles by which forms of control are possible (Checkland, 1994).

The example is fascinating on multiple levels. First, it is a brilliant example of human ingenuity, of capturing in a formal system the cross-contextual aspects of some recurring situation, and turning it into a viable solution for a problem in human life. Second, his keen observation of actual human helmsmen allowed him to note a triadic temporal structure in their behavior: the past, the present, and the future and incorporate them into a single technical system. But notice, by abstracting the relevant parameters of the situation and formalizing it into a system is in no way, now, a mark of intelligence in the mechanism, but in the human designer (and to an extent the experienced helmsmen that were being copied). Thus, when seeking real autonomy in artefacts, what needs to be recapitulated in the machine is not the superficial *form or pattern*, but the *capacity* that grounds the ability in the first place. This crucial distinction will crop up also in the distinction between automatic and autonomous. To say that a thermostat, for example, is intelligent in its behavior is surely to stretch what we mean by intelligent. To say that it is goal-directed *intrinsically,* is a mistake of the similar kind (Deacon, 2013). What the PID controller and the thermostat are, are ingenious examples of human intelligence, whose goal-directedness is a function of their embeddedness in human life: intrinsically they have neither intelligence nor goals.

3.2 Definition of Key Terms and Concepts

To launch us into the issues, what needs to be distinguished and understood in relationship to one another are the closely related terms *unmanned*, *automatic*, and *autonomous* (Williams, 2015).

Unmanned simply means that there is no human on board the vessel. The vessel itself may be autonomous, automatic, or under remote control. Unmanned is in the context of maritime vessels the goal of autonomy. The idea is that if the human element is not needed on board the ship, there will be significant reductions in the costs of maritime traffic and thus competitive advantage for shipping companies that adopt autonomy. The AAWA project (Rolls-Royce, 2016, pp. 81) cited the more efficient use of space and fuel, as well as the general optimization of shipping issuing from digitalization as the main drivers. Being unmanned further means that accidents involving unmanned ships carry no immediate human casualties. Finally, given that human error, sometimes fatigue, is a major cause of maritime accidents (Chauvin & Lardjane, 2008), if the remote operation centers and/or autonomous systems in the ships would be advanced enough, perhaps there would a reduction in maritime accidents. Of course, the conclusion that removing the human factor from the immediate ship would make human (or other) errors disappear altogether certainly does not follow – rather it recedes from the immediate situation (Ahvenjärvi, 2016).

Automatic is a closely related concept to autonomy, but there are certain distinctions which need to be fleshed out, and which will shed light on what autonomy entails. An automatic process or a system is in our definition one that has certain fixed and predetermined event flows that take it in predictable ways from one state to another. There may be some limited sensitivity to context or other environmental factors, but the defining feature is its' predetermined character and the fact that the environmental factors have been taken into account in detail beforehand. Examples of automatic processes abound in the natural and artificial worlds. A traditional machine, say for making paper clips, is automatic. The process runs along the same tracks every time, even if human ingenuity has built in mechanisms like sorting or detecting faulty items. Of course, if a sufficient amount of sorting, fault detection, or similar mechanisms are built in, the system begins to shade into *a kind of* autonomy insofar as those are tasks previously carried out by humans (Saariluoma, 2015). In the animal kingdom, some organisms seem to "run" quite rigid programs and seem slaves to their biological programming. Dennett (1990) describes the behavior of the Sphex wasp, whose reproduction strategy includes digging a burrow for eggs, finding a cricket and paralyzing it with a sting, bringing the cricket to the burrow, checking the burrow, dragging the cricket in and laying the eggs, never to return. When the behavior was studied, it was noted that if the paralyzed cricket is moved between it having been brought to the burrow's edge and the wasp going in to check the burrow, the wasp would seem to get caught in a loop. It would bring back the cricket to the edge and go in to check the burrow. This process would loop so long as the cricket

was moved after the wasp entered the burrow. Even human cognitive systems, such as perception, is to a large extent automatic. For example, certain illusions will persist no matter if we are informed as to the nature of the phenomenon (Fodor, 1985). We have no control over certain processes, or certain aspects of those processes, and the same applies for reflexes and the like. Indeed, it would hardly make sense for us to have volitional control over the faculties *as such* that tirelessly serve to build many central aspects of our experience. In cognitive science terminology, many domains of cognition are encapsulated in modules and are cognitively impenetrable (Fodor, 1985). The same goes for many of our organs: the heart goes on pumping without personal effort, the liver and the kidneys perform their functions, the pupils dilate, the hair grows, and so forth.

In the context of technological development, *automation* is the forerunner to *autonomy*. They have the similar goal of creating systems and artefacts that can manage tasks without human intervention, but autonomy seeks to expand those borders to encompass dynamic situations, and in a way that the autonomous system could manage for extended periods of time without human intervention (Endsley, 2017). Krogmann (1999) decomposed an autonomous systems' interactions with its' environment to five stages: monitoring (recognize the actual state of the world and compare it to the desired state); diagnosis (analyze the deviations from actual to desired states); plan generation (think about actions to modify the state of the world); plan selection (decide the necessary actions to reach a desired state); plan execution (take the necessary actions).

What is the role of the human on a ship? The answer seems straight-forward: human beings contribute *mental processes* to the operations of the technical system. That is, what is crucial is not that it is *a human hand* that turns the wheel that turns the rudder, but that those behaviors are instantiated in a rational manner, within a general awareness of the situation, and in terms of a goal. Explanations of this sort are perhaps best captured, as von Wright noted (2004), via practical syllogisms. Namely, the major premise of the syllogism is some envisaged end state, the minor premise relates some action as a means to that end and the conclusion is simply to use the means to reach the end. The role of human beings, especially on a ship's bridge, is to perceive, to think, to make decisions, set goals, and perform actions. All of this, in one way or another, needs to be instantiated in the artefact, or somehow circumvented, if the goal of autonomy is to be achieved. As Krogmann (1999) notes, a program that controls the behavior of a ship (for example) is not intelligent in the sense required if the software "injects" them with what to do and how to react to certain pre-specified situations. Rather what is needed is that the program has a structure that allows it to organize itself, and to learn and adapt to changing circumstances (Krogmann, 1999). This "self-organization" is akin to how a human agent "gathers" behaviors in the service of intentional action in a situation. We will address this question in more detail later. Notice however, that the previous discussion on the vision of the AAWA project (Rolls-Royce, 2016) is much more akin in spirit to automation than autonomy, although this is a question of definition rather than objective fact.

Autonomy refers to a systems capacity to act according to its own goals, percepts, internal states, and knowledge (Williams, 2015). It is, as the etymology of the word suggests, a form of self-determination. Autonomy comes from the Greek term autos (self) and nomos (law) and means *control of the self* (Bateson, 2002). To keep our eyes on the proper topic, we shall define autonomy to refer to the capacity of a causally semi-integrated system with a reasonably clearly definable inner environment to reflectively apply rational control over its' overt (and perhaps covert) behavior (Metzinger, 2017). Such a system would typically have mobility and some ways through which it can physically manipulate or negotiate its' way through the environment, and the environment itself has dynamic and unpredictable characteristic, such that the acting in it is not amenable to simple pre-programmed automation. In philosophical parlance, an autonomous system needs to have the capacity for action (Davidson, 2001; Metzinger, 2017; Taylor, 1965; von Wright 2004). We will return to the notion of action in more detail later. What is crucial to intuit at this point, is that autonomy as we are defining it leads us towards a requirement for the system to have what might in the human context be identifiable with volition, cognition, and perception. Another crucial point is that automatic or determined processes and autonomous agency, "free will", may seem like opposing concepts, but in actual practice they form a necessary unity: no volitional cognition without automatic processes, no autonomy without determination within constraints (Deacon, 2013; Wilden, 1987). The mark of autonomy is the *self-imposition of constraints* on behavior, characterized by sense-making and rational selectivity. It is a form of determination (Bunge, 1979) whose characteristic property is perhaps most generally captured by the term information.

Another way the nebulous term autonomy is relevant for our exploration is the with respect to the (quasi)autonomy of different levels of reality. Using the experience and neuroanatomy of pain as an example, Daniel Dennett (1986), discusses the irreducibility of the one to the other. It is important to note, that while Dennett (1986) certainly acknowledges the necessary connection pain has with neural networks that, for example, connect the location of the pain with the brain, he essentially treats it as an analytical question, and concludes that no mechanical explanation suffices to capture the experience of pain *as such* (see also Saariluoma, 1999). That is, the explanatory level of the person (our experiences, psychological states, intentions and so on) is autonomous with respect to the sub-personal neuroanatomical level: if we abandon the personal level and descend to its' (necessary) physical substrate, we in a very real sense abandon the sense of pain as such. And this is not a question of whether or how the two interact, they obviously do, but that the problem between the levels is *analytic* in the sense that they refer to levels that have real autonomy with respect to each other: we can't think or experience our pain away, but we can attend to other sensations, and perhaps most importantly consciously move away, eliminate, and avoid the source of pain (outside our bodies). The acknowledgement of the autonomy of the special sciences does not, however, preclude asking questions as to the reasons for their autonomy. If viewed from a systems perspective, any system save for the most

simplest, is liable to behave in ways that necessitate new conceptual tools. Music, for example, is a wholly artificial system that is nonetheless has its' own laws and theory language. Yet it is rational to ask how it comes about and having done so exhibit autonomous lawful attributes.

There is a further way in which we may exploit the term autonomy in a different context but relevant for our purposes, and which is closely tied with the preceding discussion on levels of reality. This is the autonomy of different bodies and practices of knowledge. Often this distinction is fleshed out in relationship to science, such as by Fodor (1974, 1997) with respect to sciences of the mind or by Vincenti (1990) with respect to engineering. Both argue against the proposition that the "special sciences" are nothing but applied science. That is, there is nothing within the body of knowledge that couldn't be reduced to the basic sciences: often physics and mathematics. In the context of engineering design, one might further argue against the practical sense of such a reduction, even if it were possible in principle. If one views the history of ideas metaphorically as a branching river system in which different streams of human understanding diverge and occasionally converge, we can sense that perhaps the advancement of human understanding is to some significant degree a collision of different forms of understanding (and perhaps a function of the maturity of the respective streams). To take a Kuhnian (1970) perspective, even science itself is largely a conceptual advance. It is, as is characteristic of human beings, creative but also conservative: we want new inventions and new concepts, but we do not want to lose what is valuable in the old. The collision of different communities of knowledge and cultures can yield results that transcend what could be conceived by either alone. If one is familiar with the history of cognitive science, for example, it is rather easy to see that the intellectual advance it facilitated was the result of the convergence of many different streams of thought: computer science and mathematics, psychology and the philosophy of mind, linguistics, neurosciences, to name a few (Abrahamsen & Bechtel, 2012). The point is that the mapping of different knowledge patterns unto each other is a rich source of conceptual advance, but it is certainly not a simple, obvious, or easy task to achieve. Indeed, many people have a distaste for such loose borders among conceptual structures. But this is the nature of cognitive science as an interdisciplinary endeavor. That there has always been a rich connection between computer science, cognitive science, and artificial intelligence, and the success these fields have in many ways enjoyed, speaks to the potential within the interdisciplinary approach. Insofar as autonomy, as has been implied, involves something like the instantiation of intelligence in the artefact, we will do well to keep this approach in sight.

3.3 Aspects of Autonomy

Autonomy, much like intelligence, is not some single variable for which we could devise an instrument and point it towards an artefact or an organism and expect

to find an answer (McDermott, 2007). Let's examine some ways by which attempts to get a grip on this elusive term have been tried.

3.3.1 Autonomy Levels or Scales

A common strategy to get a handle on autonomy has been to stratify it into levels both in the maritime context (Blanke, Henriques, & Bang, 2017; Lloyd's Register, 2016; Schiaretti et al., 2017a, 2017b) and more generally (Endsley, 2017; Parasuraman et al., 2000), see also (Insaurralde & Lane, 2014). Schiaretti et al. (2017a) have argued that assigning an artefact's level of autonomy to a single variable on a leveled hierarchy is not fine-grained enough, given that there are several subsystems and classes of functions that together are thought to yield autonomy, but which may have different levels of development in different embodiments, making simple arithmetic possibly misleading (see also Williams, 2015). This is likely to be true, but we shall bracket it from consideration for now, and approach the relatively simple autonomy levels outlined by Blanke et al. (2017).

Table 1 Autonomy Levels in the Maritime Context according to Blanke et al., (2017)

Description	**Operator role**
AL 0: Manual steering. Steering controls or set points for course, etc. are operated manually.	The operator is on board or performs remote control via radio link.
AL 1: Decision-support on board. Automatic steering of course and speed in accordance with the references and route plan given. The course and speed are measured by sensors on board.	The operator inserts the route in the form of "waypoints" and the desired speed. The operator monitors and changes the course and speed, if necessary.
AL 2: On-board or shore-based decision support. Steering of route through a sequence of desired positions. The route is calculated so as to observe a wanted plan. An external system is capable of uploading a new route plan.	Monitoring operation and surroundings. Changing course and speed if a situation necessitates this. Proposals for interventions can be given by algorithms.
AL 3: Execution with human being who monitors and approves. Navigation decisions are proposed by the system based on sensor information from the vessel and its surroundings.	Monitoring the system's function and approving actions before they are executed.
AL 4: Execution with human being who monitors and can intervene. Decisions on navigation and operational actions are calculated by the system which executes what has been calculated according to the operator's approval.	An operator monitors the system's functioning and intervenes if considered necessary. Monitoring can be shore-based.

(continued)

Table 2 Autonomy Levels in the Maritime Context according to Blanke et al., (2017) (continued)

AL 5: Monitored autonomy. Overall decisions on navigation and operation are calculated by the system. The consequences and risks are countered insofar as possible. Sensors detect relevant elements in the surroundings and the system interprets the situation. The system calculates its own actions and performs these. The operator is contacted in case of uncertainty about the interpretation of the situation.	The system executes the actions calculated by itself. The operator is contacted unless the system is very certain of its interpretation of the surroundings and of its own condition and of the thus calculated actions. Overall goals have been determined by an operator. Monitoring may be shore-based.
AL 6: Full autonomy. Overall decisions on navigation and operation are calculated by the system. Consequences and risks are calculated. The system acts based on its analyses and calculations of its own capability and the surroundings' reaction. Knowledge about the surroundings and previous and typical events are included at a "machine intelligent" level.	The system makes its own decisions and decides on its own actions. Calculations of own capability and prediction of surrounding traffic's expected reaction. The operator is involved in decisions if the system is uncertain. Overall goals may have been established by the system. Shore-based monitoring.

What we are ultimately interested in in this thesis is the level of full autonomy, and its' possibility from current technological solutions from a foundational perspective. This level comes surprisingly fast in the hierarchy above: already on level four there is an expectation that the technical system has in principle the capacity to make and implement decisions, and the human is already receding from the immediate artefact to a supervisory role. It is true, however, as Schiaretti et al. (2017a) argued, that different subsystems in an autonomous ship are likely to have and develop at different speeds in terms of autonomous capacity (see also Williams, 2015). Furthermore, different subsystems may have different levels of autonomy at different times, as when the task environment becomes too challenging for the system to handle, and it should relinquish autonomy to human operators in shore control centers. How the system will be built such that it can identify those situations is an important question by itself. At any rate, what the scale above combines in a rough way are on the one hand the sharing of duties, or more precisely when the human operator should assume control of the vessel, and the corresponding capacity of the technical system to handle certain tasks and/or situations. But giving a categorical answer seems difficult. For example, certain routine tasks to do with collision avoidance are quite different in terms of difficulty when the potential encounter is between two ships in the open ocean as opposed to a congested shipping lane with several ships (Rolls-Royce, 2016). Thus, the capacity of a system to perform autonomously should be evaluated against multiple dimensions simultaneously and with respect to rigorously specified scenarios. We should ask, for example:

For technical system S in scenario X: *characterized by environmental complexity c(E) and task complexity c(T), where c(E) refers to the unpredictability and number of relevant elements and variables in the environment with respect to the task, c(T) refers to the decision-tree complexity that results from the available moves or actions the successful completion of the task requires as evaluated against c(E),*

the system is capable of autonomy only if the system has the sufficient capacity in autonomy dimensions d_1 – d_n (for example goals, sensing, perceiving, apperceiving, acting, decision-making, etc.) which are required to complete the task in accordance to criteria c_1 - c_n (for example safety, efficiency, rationality, higher-order goals and constraints, etc.).

Thus, autonomy is a relative term that is subject to variation (in time and across contexts) based on a complex interplay between the environment, the task, and the system's capacities (Williams, 2015). The benchmark and reference point for the capacities is, of course, the experienced human seafarer and the system in which he or she is embedded in. As it is the human who is being replaced from the immediate vessel, it is rational for us to ask *what* capacities underlie the human ability to deal with particular scenarios, and *how* those capacities function. For example, it is only as a reference number that the complexity of chess is brought up in analyses of human thinking as they play the game, far more important and interesting is how that potential complexity is narrowed down in mental representations through apperception, for example (Saariluoma, 1992). That is, in human affairs, it is the contents of our experience that guide behavior and action, but that content is only partly available in the stimulus "as such", and indeed in chess the colors and placement of the pieces only give a snapshot for the player, the point of the game being to test the player's ability to plan, imagine, and apperceive to a significant degree "in their heads". Furthermore, it is quite a different exercise to take one form of human activity, say the game of chess, and attempt to coax a mechanical system to perform well in it, than to attempt to recreate the general capacity of humans to adapt and learn some rule-based form of life such as chess in the machine (Lake et al., 2017). Such general ability is still, according to expert opinion (Müller & Bostrom, 2016), at least decades in the future.

The question of how human beings manage the tasks they do is of course a difficult one to answer even if the target were only human capacities, let alone here where as one descends down levels of analysis with regards to a single capacity, one is likely to discover a disconnect between the ways human beings and artefacts fundamentally operate[2]. In addition, the human operator is not an isolated individual but embedded in a social practice which extends the analytic space outside the individual to the crew, other ships, and so on. There is no guarantee that a reconciliation is possible in the constitutive sense, although a perfor-

[2] Opinions as to this point vary and such a "descent" is tied to one's conceptual assumptions. See for example Shanker (1998) for a discussion on the "continuum picture" of human mental ability vis-à-vis information processing in computers. We will address this in detail later.

mance-based approach is possible, and indeed typical and traditional for artificial intelligence (see for example Turing, 1950). Should we, however, find that the constitutive gap is unbridgeable, and that the constitution makes a difference with respect to the system being able to reach a certain level of performance or satisfy some criteria, then that would call for a re-evaluation and re-orientation of the principles that guide the attempt. We shall not anticipate this, but merely flag it as a possibility. It is nonetheless important to recognize at which level of abstraction the requirements translated from human capacities should be provided for engineering design. For immediate practical purposes an intermediary language that strikes the right balance between generality and detail is likely to be most useful for engineers. For longer term purposes both with respect to cognitive science and artificial intelligence, the deep questions as to the differences and similarities between machines and organisms, the questions of what mentality is, all unresolved philosophical and practical questions, need to be addressed. We will attempt in this thesis to take a stab at both directions.

This brings us to a crucial distinction between *constitutive* and *ascriptive* autonomy (Rohde & Stewart, 2008). Turing's approach, with regards to intelligence, which can still be seen to be a guiding method in AI, was the latter. Turing (1950) wanted explicitly to bypass the problem of (other) minds in the sense of looking for constitutive characteristics of the mind or processes by which it emerges or the physical properties of the substrate upon which it is instantiated. His method, the imitation game, was to essentially allow for a level playing field between machine and man in the evaluation of intelligence given the assumption that the ascription of intelligence even to other humans is essentially a leap of faith itself. It could be said, that Turing (1950), quite rightly, felt that our ascription of intelligence to a system depended, or was influenced by, the knowledge we have of its' constitution. In other words, one might say that we would be biased in ascribing, if not intelligence, then at least mentality to a computer given the deep connection between life and mind that we generally intuit, see Thompson (2010) for a discussion on this concept. It should be said, in this connection, that Turing (1950) assumed that this manner or thinking was something that would be subject to change as technology progressed and the adoption of intelligent machines would become more commonplace. It is indeed a clear possibility that insofar as technology obtains human-like characteristic such as speech and naturalistic body movements, we humans are liable to begin ascribing mentality to them. But it could be argued that it is precisely at this juncture that our understanding of the constitutive aspects of mentality or intelligence, insofar as they connect with autonomy, will become a moral and pragmatic imperative. More interesting for our purposes, however, is the question whether even the *performance* of smart machines can cross some threshold without our understanding of the *constitutive* aspects of mentality.

3.3.2 Dimensions of Autonomy

As has been indicated, there are distinct problems with assigning autonomy as a single property of a system. As Williams (2015) notes, no system – human or otherwise – is completely autonomous with respect to its environment or even its' own subsystems. Autonomy is a term that used in various ways, with various meanings and senses. What can generally be accepted is that in the context of maritime vessels, autonomy refers mostly to the attempt of replacing what was previously human action with artefact action. An autonomous system refers in this sense to technologies that have the capacity to perform tasks that previously required the higher cognitive abilities associated with human thinking (Saariluoma, 2015). The way towards this goal will likely proceed by identifying specific tasks and problems in the maritime context as it relates to the decision-making and implementation on the ship's bridge, and seeking technical solutions to them one by one. Our task is to map out the problem-space on a general level. Thus, we want to know what are the general dimensions of human cognitive ability that correspond to the tasks and problems which the technical solutions are seeking to replace. Williams (2015, 54) summarizes the key dimensions of autonomy for technical systems as follows.

Table 3: Key dimension of Autonomy (Williams, 2015)

Autonomy dimension	Definition
Goals	An autonomous agent has goals that drive its behaviour.
Sensing	An autonomous agent senses both its internal state and the external world by taking in information (e.g., electromagnetic waves, sound waves).
Interpreting	An autonomous agent interprets information by translating raw inputs into a form usable for decision making.
Rationalising	An autonomous agent rationalises information against its current internal state, external environment, and goals using a defined logic (e.g. optimisation, random search, heuristic search), and generates courses of action to meet goals.
Decision making	An autonomous agent selects courses of action to meet its goals.
Evaluating	An autonomous agent evaluates the consequences of its actions in reference to goals and external constraints.
Adapting	An autonomous agent adapts its internal state and functions of sensing, interpreting, rationalising, decision making, and evaluating to improve its goal attainment.

Saariluoma (2015) includes in the list of human cognitive abilities pertinent for autonomy processes such as categorization, concept formation, learning, judgement and inference, decision-making, and problem-solving. Such a dimensioned account clearly gives us more grasp over what autonomy entails from the cognitive and AI perspectives. These accord well with the general architecture of the ANS system described before (Rolls-Royce, 2016). It should be noted, how-

ever, that we are now simply projecting the necessary dimensions on the architecture, without any real ground for expecting to find such "cognitive" dimensions in the artefact. But if the ship is to be more than automatic, and thus necessarily under human supervision, there should be some principled projection that takes place from these dimensions or capacities onto the ANS system or some of its modules.

To deepen our understanding of what such autonomy dimensions may entail, let's turn our discussion towards a general assessment of what autonomous artefacts are up against from the perspective of the environment as evaluated against perception and cognition.

3.4 Reality From the Perspective of Perception and Action

One justification for the attempt of creating unmanned ships is, in addition to economic or ecological considerations, that it could result in an overall safer maritime ecosystem. First, any disaster involving only unmanned vessels would cause no human casualties, and second, since human error is a major cause of maritime disasters (Chauvin & Lardjane, 2008), the hope is that perhaps a well-enough designed autonomous ship could avoid such disasters. But whereas human error, resulting perhaps from fatigue, incorrect situational awareness, incomplete or false information, or the limits of perceptual and cognitive systems, is a well-studied phenomenon (Endsley, 1995, 2015), the types of errors resulting from the decisions made by an artificial intelligence are less well understood[3]. This would be doubly the case if the ship's intelligence would be the result of a neural network, whose complexity and inscrutability rises as a function of their effectiveness. Indeed, according to MIT professor Patrick Henry Winston, "no one knows what the neural nets are doing", and "a cottage industry has emerged where researchers try to fool neural nets" (Winston, 2016). See figure below for an illustrative example.

[3] Or perhaps more accurately, they are more *alien* to us. Take perception for instance. Although machines have surpassed human accuracy in image classification tasks (Shoham, Perrault, Brynjolfsson, & Clark, 2017), the types of errors they make are rather telling. A small ant on a blade of grass, or a human face put through effects gives humans no trouble – but a neural network simply can't recognize the forms under these unfamiliar distortions. See also Nguyen, Yosinski, & Clune (2015).

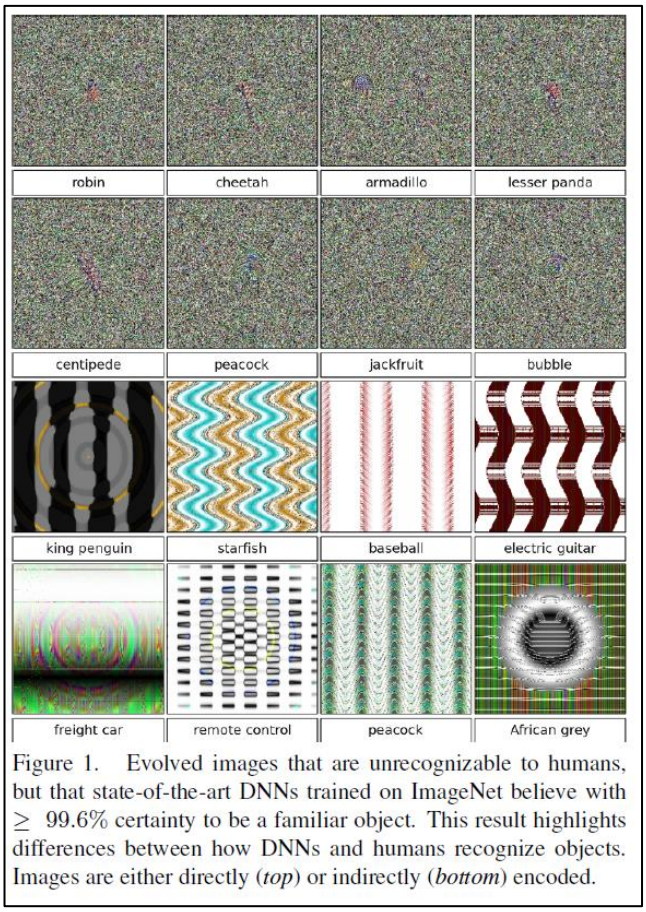


Figure 1. Evolved images that are unrecognizable to humans, but that state-of-the-art DNNs trained on ImageNet believe with $\geq$ 99.6% certainty to be a familiar object. This result highlights differences between how DNNs and humans recognize objects. Images are either directly (*top*) or indirectly (*bottom*) encoded.

Figure 5 High confidence predictions from DNNs in images unrecognizable to humans. From (Nguyen, Yosinski, & Clune, 2015).

On the other hand, control algorithms of a simple deterministic nature (automatic in our language) are unlikely to be effective in a complex-enough environment, such as the sea, and more sophisticated systems, even traditional software, can easily become inscrutable and thus may carry the possibility of disastrous hidden error. It would be wrong to assume that the automation process of ships moving and interacting with the environment would be solvable from the paradigm of, say, factory automation where the process can be well-defined and broken into distinct phases, and the environment closely controlled. Indeed, the machines operate often in complete darkness and humans are not allowed into the factory floor. Nor would such be anything like autonomy in the sense we are investigating. The devil, as is often said, is in the details, but it may also be within the nature of the system. Any system used to implement autonomous decision-making and steering on a ship needs to be scrutinized down to its' minute details, and its' decision-making processes made explicit, but since no such system exists in implementation, we will approach the question from a broader angle, and in the process re-state age-old foundational questions that lay at the roots of AI (Artificial Intelligence).

From the perspective of environment and action, what are autonomous technical artefacts up against, generally speaking? It can be argued that a central

aspect of human cognition that sets it apart from machines, and to some extent other animals, is its' selectivity (Saariluoma, 1992; Saariluoma & Rauterberg, 2016). This selection is achieved against a reality that is in principle inexhaustible in terms of *interpretations* (Peterson, 2013; Saariluoma, 1992). Consider the example before of high-confidence interpretations by neural nets. The nets are picking up *some* pattern in the image but they clearly have no understanding of what they are seeing from a human perspective. They do not *see.* Our perception is tuned to a level of reality roughly corresponding to our biological inheritance, that is graspable objects, other beings, places of shelter, sources of food, often referred to as affordances (Gibson, 1966). We easily take for granted how much we contribute in terms of innate organizing principles, for example causality (Lake et al., 2017; Marcus, 2018).

But the patterns which we can identify are to some extent malleable, and in certain cases seemingly arbitrary, such as chess (Saariluoma 1992, 1999). We can't perceptually tune in to the patterns active at the atomic, chemical, or cosmic scales, but we can *think* about and apperceive such patterns and build instruments that can perceive or seek to establish if such patterns exist in reality. Perception is intimately tied up with action (goal-directed behavior). But action complexifies the picture even further. For example, given twenty possible moves to be executed over a series of ten actions the exhaustive examination of all possible combinations would correspond to a number with thirteen zeroes (Saariluoma 1992). This is nowhere near the processing capacity of human beings, and yet such series of actions are hardly uncommon. What this means is that human information processing is heavily selective and can capture larger swathes of reality within manageable limits, in service of actions and goals. Indeed, it is our very capacity to perceive the world in different ways depending on our goals that, so far, marks man apart from machine.

The relevance from the perspective of the design of autonomous artefacts is clear. However human beings manage this feat of selection is likely to be informative for the design of intelligent artefacts. But consider a classic benchmark in artificial intelligence, when Deep Blue won against Garry Kasparov in chess. From the perspective of the task (of playing chess) the result is undeniable. But from the perspective of AI, the solution to the problem was not even close to the methods of human problem-solving. The way human beings capture relevant patterns in a board of chess (Saariluoma, 1992), is rather different than the methods employed by Deep Blue. This is not to deny the feat as remarkable. But one should bear in mind that chess is in terms of rules is a constrained space that can be absolutely determined. The relatively few moves and pieces, still, yield a complexity (of variations) for the game on the order of 10^{120} (Shannon, 1950). What about close encounters among multiple ships in congested settings? We shall not attempt a precise calculation here, if such is even possible in principle (which is a part of the problem). But if we account for the ships possible movements through six degrees of freedom and hypothesize an encounter between say four ships, some of which are autonomous, others human experts, and some amateur sailors that may (or may not) follow the maritime rules of the road, COLREGS,

themselves somewhat ambiguous, we can quickly intuit that the problem space from a machine perspective tends towards the complex. Complexity is almost by definition outside what can be decided by calculations, and one is therefore tempted to say outside the capacity of computers, insofar as the central requirement for algorithms is unambiguity. A fundamental concept for computer science is that a set of rules (the program) *unambiguously* specifies certain processes which can be carried out by a machine processor built in such a way as to accept these rules as instructions determining its operations (Boden 1987, 7).

The central issue at stake is not that human beings have more "calculating power" "in their brains", but that the immense selectivity of human thought is qualitatively different from a formal model by which a machine must operate. Human beings, especially experts, can quickly narrow a situation to a few *relevant* patterns, and *act* based on that (Saariluoma, 1992, 1997). The fact that we can tune in to a system as artificial as chess, speaks to the flexibility and hence non-programmatic quality of our thinking, not to mention our capacity to *create* such arbitrary games. The human capacity to *create* and *follow* rules needs to be appreciated and understood as a crucial difference between us and machines (Shanker, 1998). Post-hoc, we may identify those patterns and (in principle) construct programs built around them, yielding software somewhat capable of dealing with the *exact* same situation. But what if the situation changes, and how much can it change for the same program to still yield satisfactory results? This is an empirical question to be settled on a case by case basis, but the relevance for our larger discussion is that the capacity that should be instantiated in the machine is not a rigid behavior pattern only, which is automation in our definition, but a selection and an adaptive application of a general pattern in a rational manner. The latter is the mark of autonomy, and the distinction between automatic (rigid, pre-specified) and autonomous – linked as though they may be – becomes apparent once again.

3.5 Summary

The attempt of this section is to orient ourselves around the topic of autonomy in general, and in the context of maritime vessels. We began with a relatively pragmatic and descriptive assessment of the vision for autonomous ships, as laid out in the AAWA project (Rolls-Royce, 2016). The vision for the project was seen as relatively modest and realistic given current state of technology. Our discussion of the themes of autonomy in general lead us however to consider what the limitations of a grander vision of autonomy might entail given that it may feed into our desire to understand the way human-work interaction or sharing of duties should be laid out in the foreseeable future.

The main message was to establish a link between cognitive processes in man and autonomy in the sense we define the term: as the capacity of a system to self-determine its' behaviors in accordance with goals and the state of the environment. Insofar as autonomy entails the replacement of human thinking from

the technical system, it becomes inextricably linked with artificial intelligence. As soon as this move is introduced, it opens up many difficult problems in cognitive science and artificial intelligence as they pertain to the unique characteristics of the mental.

4 ARTIFICIAL INTELLIGENCE

In the last section, we concluded that autonomy proper, as opposed to merely unmanned and automatic, requires the instantiation of certain key dimensions in the artefact. These, in turn, have a distinct cognitive, mental, and intentional characteristics. Thus, proper autonomy requires something like artificial intelligence. The purpose of this section is to examine foundational issues around machines, computers, and artificial intelligence to get a better grasp on the operating principles and concepts upon which AI is building.

4.1 General Introduction

> The original question, 'Can machines think? ' I believe to be too meaningless to deserve discussion. (Turing 1950, 442)

These lines are from a seminal paper on artificial intelligence, Alan Turing's "Computing machinery and intelligence" from 1950. Yet Turing continues,

> Nevertheless, I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted. I believe further that no useful purpose is served by concealing these beliefs. The popular view that scientists proceed inexorably from well- established fact to well-established fact, never being influenced by any unproved conjecture, is quite mistaken. Provided it is made clear which are proved facts and which are conjectures, no harm can result. Conjectures are of great importance since they suggest useful lines of research. (Turing 1950, 442)

According to Poole and Mackworth (2010) the field of artificial intelligence is "the synthesis and analysis of computational agents that act intelligently". Agent here simply means an actor, someone or something which does something in an environment, but such that its actions can be characterized as intelligent: its' behavior is appropriate with respect to the environment and its' goals, it learns from experience and adjusts its goals and behavior flexibly. The computational part means that AI builds on the fundamental operating principles of computers, while also expanding those principles to include human information-processing, much like the mainstream of cognitive science (Thagard, 2005). The engineering goal of AI is to design and build intelligent systems for human purposes (Boden, 1987). The scientific goal of AI is to be a kind of science of intelligence *in general* (Boden, 1990): to understand the principles that make intelligent behavior possible in natural or artificial systems (Poole & Mackworth, 2010). As Drew McDermott (2007) put it, AI and cognitive science get their ideas from fields like computer science, psychology, linguistics, neuroscience, and philosophy and give back systems and models of information-processing that indicate whether those ideas work. In a

way, you could say that AI is in an interesting sense the empirical part of cognitive science, at least insofar as we refer to the (traditional) part of cognitive science that posits computations and representations as the fundamental operating principles of mind (Frankish & Ramsay, 2012; Newell & Simon, 1961; Thagard, 2005). To grossly overgeneralize to make a point, the fundamental approach of cognitive science, if it to be taken as more than a metaphor or method of modeling, is at stake in AI – or at least the limitations thereof.

The posing of the problem of intelligence in the context of actually attempting to achieve it by modeling, testing, and implementing it in computational systems lead to an exciting time in the science of mind and intelligence (Abrahamsen & Bechtel, 2012). Concurrently with AI a new discipline called cognitive science took the same core ideas of information-processing and computations towards understanding mentality in humans.

The history of AI that followed is one of high ambition and exaggerated expectation, followed by disappointment and increased appreciation (in a generation) of the magnitude and difficulty of creating human (or even animal) level intelligence in a machine (Dreyfus, 2012). But the wheel turns, and it seems that *this time*, we have the right ideas and tools to make for that final last mile towards genuine AI. Anyone following the press and media over the last couple of years can't have failed to note that we seem to be in such an age once again (Lewis-Kraus, 2016). Being in this age feeds into many different attempts, importantly for us into the attempt at creating autonomous technology, such as unmanned ships. There are three advances which have led and fueled current enthusiasm over AI: the availability of big data; improved machine learning approaches and algorithms; and more powerful computers (National Science and Technology Council, 2016).

The roots of AI can in part be traced to a seminal paper by Warren McCulloch and Walter Pitts (1990), which introduced the idea of formal neurons based on the principles of Boolean models[4]. This, in turn, directly influenced von Neumann and subsequently became the basis for the logical design of digital computers (Boden, 1990). Indeed, the common birthright of both so-called traditional symbol-processing models and connectionistic (neural network) models in AI are to be found in the McCulloch and Pitts paper (Boden, 1990).

The term artificial intelligence was coined by John McCarthy in a proposal for a research project to take place during the summer of 1956 at Dartmouth College in New Hampshire, USA (McCarthy, Minsky, Rochester, & Shannon, 1955). The study was "to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it." Some 60 years later, this conjecture remains a conjecture and the ambitions set forth by McCarthy and his colleagues remain unfulfilled. Indeed, McCarthy (2007) later noted in a review of

[4] It is less well known however, that the roots of neural nets trace back further to Nicolas Rashevsky, the founder of the journal, Bulletin of Mathematical Biophysics, in which the original article was published (Cull, 2007).

problems towards genuine AI how common-sense informatic situations characterized by ambiguity and vagueness presented the greatest challenge for AI. A perhaps reasonable modification of the (satisfyingly, one might add) bold statement given what has been achieved might go something like: "*some* aspects of learning or intelligence can be so precisely described that a machine can be made to *simulate* them". The question that looms is which aspects, and therefore what are the limits of machines? For even as our understanding of the problems of AI have been brought into sharper focus (see McCarthy, 2007), the fundamental operating principles of computers have not changed.

4.2 Shared history of AI and Cognitive Science

It may not be obvious to all readers that cognitive science, computer science, and artificial intelligence all share a common and intertwined history. The classical computational-representational view of cognition is derived from the pioneering work on computers by Turing and Von Neumann (Fodor & Pylyshyn, 1988). A.I based on rule-governed manipulation of formal symbols as exemplified in classical computers is often (rather derogatorily) referred as GOFAI, short for good old-fashioned artificial intelligence (Boden, 1990). According to Boden (1990) both GOFAI and connectionism (or neural nets) trace their lineage to a seminal 1943 paper by Warren McCulloch and Walter Pitts (McCulloch & Pitts, 1990) which synthesized ideas from the neurons of the brain, the computable numbers of Turing, and the work on propositional calculus by Russell and Whitehead (Boden 1990, 3). The significance is that both GOFAI and connectionism are theoretically grounded in Turing's paper on computable numbers, which defined computation as the formal manipulation of symbols, by an application of formal rules (Boden 1990, 4). While perhaps not representative of cognitive science today, fractured and diverse as it is, in the early days AI was seen as "the glue that would bind together such diverse fields as psychology, linguistics, anthropology, neuroscience, and philosophy under the umbrella of cognitive science" (Shanker, 1998).

Cognitive science arose as a response to two things: behaviorism as a paradigm in psychology and computers as an invention. Cognitive science reacted against the mindlessness of behaviorism, and took the principles behind computers as the implementation method for mind: logic, data, and algorithms that were realizable on any physical platform. Generally speaking, cognitive science attempts to explain our mental life by positing *mental representations that have semantic contents and computational processes that operate on them* (Frankish & Ramsay 2012, 34). By seeing the mind as a computational system, an idea derived from the structure of Turing and Von Neumann machines (Fodor & Pylyshyn, 1988) cognitive science is ready to equivocate, on a general level, computers and brains in the sense that both of them are *systems that process information*. A bold example

of this idea is found in Newell and Simon (1976): for them the mind is a computational system; the brain literally performs computations; and these are identical to those that could occur in computers.

It is important to note, that cognitive science is attempting to penetrate under the surface of what we would call our folk-psychological notions of say, cats, dogs, or mathematical equations. It is precisely to do this that the field posits mental representations and information processes under the "surface" of our experience, since it is at that level that we may discover the impersonal, non-idiosyncratic, or in other words scientific processes that underlie our experiences (Von Eckardt, 2012). This level of analysis should then be able to do two things: first, connect with our personal experiences in some principled manner, and two, be analyzable as the atomic structure of mental life which operates by simple, perhaps even mechanical, rules that lack the intelligence which we attribute to ourselves as a whole (Shanker, 1998).

The way AI fits into the picture (as a theory of mind) is by positing simple, mechanical steps which lack intelligence in themselves, combining them into modules, programs, and software which each stand for the mind-brain at some level of abstraction. Thus, the steps in the program become the interface between the mind and the brain, and the program as a whole becomes the mind, or at least a theory of its architecture. What AI can then offer is a precise description of the sub-personal processes that cognitive science is seeking to understand —given, of course, that the assumptions are true.

The basic assumption shared by both AI and cognitive science is this: once you drill down beneath the surface of everyday folk-psychological concepts and indeed our experiences, you will find vast and fast moving arrays of simple, mechanical computations that are in themselves meaningless and unintelligent. According to this view, mental life is characterized by a sort of continuum which extends downwards into simple mechanisms and which we share among other animals (and machines), ours being simply a more sophisticated form of the same fundamental principles. According to Shanker (1998, 50) in this view the following statements would share the same lineage: the thermostat clicked on; the leaves of the plant turned towards the sun; the pigeon pecked the yellow key in order to get a food pellet; Kanzi pressed the drink lexigram in order to get a drink; S has learned to play the Toccata and Fugue in D-minor. They are all similar in kind, and only distinguished by their ascending mechanical complexity. In an illustrative example of the continuum picture, it has been said (Shanker 1998, 135) that "the psychologists goal must be to deliver a complete (literally gapless) description of the *causal* connections that govern the total course of intellectual and/or motor processes in problem solving." Thus, the gaps that human beings exhibit as they describe their own thought processes in verbal protocols are taken to be evidence for the hidden, underlying, subconscious processes that underlie thought. Thus,"the epistemological framework underpinning the mechanist thesis is the premise that there is a gap between input and action in the exercise of an ability which must be bridged by a series of internal operations" (Shanker 1998, 59). In other words, "*if* the agents brain is seen as some sort of a computational

information-processing device, *then* the fragmentary evidence presented in the subjects verbal protocol *must* be treated as the conscious elements of a mental program whose underlying operations are pre-conscious, ie inaccesible to introspection, but inferrable from the corresponding steps in a computer program that is mapped onto the protocol." (Shanker 1998, 71, emphasis added). Thus Turing's question, "can machines think?" actually becomes whether "thought can be mechanically explained". To the extent cognitive science follows this assumption, it will eventually have to deal with how mechanical explanations can be combined in such a way that it can explain (not explain away) all facets of mental life, and ground them in a robust manner. The emphasis on "pre-conscious computations and information processes" almost by definition turns conscious intentional mental life into a *problem* to be explained by the methods outlined before. This is a serious handicap, given that it is perhaps the central aspect of (mental life). As Fodor (1985) noted, it may be that whenever semantic (as opposed to syntactic) or global (whole integrated information) features of mental processes appear, the limits of Turing-style computational rationality become apparent – and what lies beyond those limits is not a problem, but a mystery, given current state of understanding.

4.3 The Computer

It can be said without much speculation that the most viable artificial system for implementing intelligent or quasi-intelligent behavior is the computer. We will focus on digital computers, given that they are the kinds of computers usually used and most widespread.

Computer science is, as Aho and Ullman (1994) put it, the *mechanization of abstraction:* the attempt to model certain information-processing problems and devising mechanizable techniques to solve them. While computers are often said to process information, it is perhaps more accurate to approach their operating principles from the perspective of data and algorithms. The term information is used in many senses and it is not exactly wrong depending on context to say that the computer processes information, but it should be noted that information insofar as it connects with meaning is better reserved for systems that can without a doubt be said to have capacity for meaning, namely human beings in this context. The processes of the computer are information-processing only to the extent that the process and results thereof make sense within the context of human affairs: a computer churning out calculations in a dead universe exhibits only orderly physical change as specified by its' mechanisms. This is more than mere semantic conservatism (or atavism as one might have it). The operating principles and essential nature of the computer can get lost in the fog if certain ontological commitments are not adhered to. As we will see, mechanization, abstraction, data, and algorithms are extremely useful terms to get a handle on what the computer is and what it does. Furthermore, the mold into which intelligence is

to fit, if it is to fit in a computer, may be leaving out parts that are essential for the whole endeavor.

The particular kind of machine used in AI is the digital computer. The electric operations of the computer are directed by a binary machine code that is further abstracted into algorithms and functions in programming languages, which help humans operate and write programs for computers. The computer only 'understands' machine language - programming languages are collections of shorthands and abstractions that have been developed for human convenience. An algorithm consists of a set of rules, which are all of the same, trivial, complexity which together yield a specific output from a specific input (Dym & Brown, 2012; Shanker, 1998). Programs tell computers what to do: A fundamental concept for computer science is that a set of rules (the program) unambiguously specifies certain processes which can be carried out by a machine processor built in such a way as to accept these rules as instructions determining its operations (Boden 1987, 7). Thus, computers exhibit a hierarchical and modular structure: the foundational electro-logical operations are both interpreted and directed by the programs, which in turn have different levels of abstraction. The computer hardware is also modular, with different components taking on various roles and functions in the operations of the whole: the hard-drives store long-term data, RAM acts as the short-term "memory", the central processor directs and computes the various strands of computations, etc. Computers are not exactly calculators, rather they manipulate symbols. A symbol is an inherently meaningless cipher that becomes meaningful by having meaning assigned to it by a user (Boden 1987). The differences among computers, say a Mac and a PC, lie in the specifics of the machine language, architecture, and hardware. They do not however differ in their fundamental operating principles.

4.3.1 The Mechanization of Abstraction

The jig is essentially up from the perspective of strong AI (Searle 1980, 1984, 1990) already with our first definition, the mechanization of abstraction. Many would cling to the word mechanic and might argue against the proposal that human thought predicated on biological functioning is essentially non-mechanistic at least to a considerable degree. We will focus rather on the term abstraction, because something like mechanisms are clearly found in biological systems, and furthermore mechanism need not be conceived of as exactly *mechanistic* or deterministic (Bunge, 2004). According to Bunge, a mechanism can be conceived of simply as that which makes some system "tick" be it an economic, social, mental, or metabolic. Abstraction, on the other hand, reveals more clearly the role of human beings in the operations. It is important to understand that the weakness and power of computations are one and the same: they are achieved by a process of abstraction that leaves only the algorithmic processes in place (Saariluoma & Rautenberg, 2015) and jettisons the contents. To the extent a process of human thought can be described precisely as a token-manipulation process, it can in theory be simulated on a computer. The computer does not understand the meaning

at any stage of the process, but upon human evaluation we can judge the end-result to be correct or the steps in the computation to be plausible as a psychological model. The problem is, of course, that human thought can't be fully described as a formal system, because a formal system is predicated on the idea that given a set of axioms, one can "reason" within the formal system purely mechanically, and no intelligence or understanding is needed (Shanker, 1998). Even if you would have a multitude of different formal systems (frames or scripts) for different situations, you would still need some way of deciding which one is relevant for a given occasion, and such a procedure would imply infinite regress if your only method for the decision are more formal rules (Dreyfus, 1972). Thus, the question of relevance and even truth always seems to move at right angles in regards to formal systems, it seems to imply "a view from the outside" (Penrose, 1990). Thus, the question is whether the ideas of AI (and cognitivism) can be used to *ground* mental life as a whole and be derived from those premises? Or are formal systems just a footnote to the list of capacities of human minds – irrespective of whatever intrinsic properties they may have outside of human psychology?

The powerful advantage (powerful enough to be a shortcoming in some contexts) of attempting to instantiate intelligent behavior in computers is the absolute rigor and lack of vagueness demanded by the platform (Dym & Brown, 2012). The set of rules must be unambiguously defined for the machine processor. Thus, you can't really fool your way with computers, a thorough understanding is needed of the requirements of a particular class of required behavior, and also the constraints imposed by the computer by its' very operating principles. Another powerful advantage (and shortcoming) of using the computer is that forces "banal information", information we take for granted phenomenologically, to the surface (Dennett, 1990). This is due to the simple reason that the computer starts at-zero, it has no knowledge of the world and every required item must be somehow impressed upon the computer either by a prescient programmer or a learning algorithm of some sort. It is truly a tabula rasa. This is far from being a trivial problem, perhaps easily disguised as such due to the seeming ease with which we make us of "banal information" in certain contexts. We easily take abstractions such as events, actions, objects, and properties for granted (Veres, Molnar, Lincoln, & Morice, 2011) and wrongly assume that such boundaries are given by the environment or the stimuli thereof. The problem of how goals and context narrow and define our perceptual world and guides our action, comes home to roost in AI, precisely because nothing can be taken for granted without inviting trouble down the line. In a recent critique (of deep learning, a technique in AI) Gary Marcus (2018) pointed out how even our most sophisticated pattern recognition software has trouble generalizing beyond the dataset on which they've been trained. But with a human being, we can be at least in some sense sure that if they are given some instructions, the mistakes they would make would more likely be a case of fatigue or misguided attention, rather than some completely for us alien mistake such as not recognizing an object as a person only because the angle is such that we have never seen it before. The analytical problem is that

computers can't make mistakes due to their deterministic nature. Computer programs may have bugs, unexpected results issuing from perhaps mistakes by the programmers or having overlooked some aspects of the context of the computations, but it is not strictly speaking correct to speak of the machine as making a mistake. This presupposes a flexibility that is afforded to human mental properties, such as rule-following and mental representations (Shanker, 1998).

To put the problem succinctly, a computer or a computer program do not, as a formal-syntactic system, determine the extension (the contents) of the symbols over which the operations are defined (Shagrir, 2005). This requires an observer who can assign meaning to the symbols. This assignment can't be made from within the formal-syntactic system (Rosen, 1999), as there are no system internal methods of escaping "the Chinese room" (Searle, 1980).

4.3.1.1 John Searle, the Chinese Room, and Hubert Dreyfus

Along with Turing's (1950) imitation game, perhaps the most famous thought experiment in AI is John Searle's Chinese room (1980, 1990). Searle sought to criticize what he called the Strong AI thesis. In contrast to Weak AI, which was the philosophically relatively unproblematic task of seeking understanding of mental processes by modeling them on a computer, or building quasi-intelligent software to fulfill human needs, Strong AI proceeded from the thesis that there was literally no difference between a computer and a brain. Consequently, the right program backed with sufficient computing power would be literally a mind, no different in essence from the human mind. Searle's argument was, however, not based on any limitations of algorithms or computing power, but on the conceptual difference between syntax and semantics. Syntax refers to the form and the rules that govern a language. Semantics refers to the study of the meanings and systems thereof corresponding to language. In the Chinese Room, a human being without knowledge or understanding of the Chinese language sits before a bowl of Chinese symbols and a complicated rulebook. From outside the room, native Chinese speakers are asking questions by passing Chinese symbols into the room. The task of the person in the room is to match the incoming symbols, based on the rulebook, to other symbols in the bowl in front of him. Supposing the rulebook is sufficiently complex and thorough, and the Chinese people outside the room are satisfied with the answers, are we now in the position to say that the person in the room understands Chinese? Searle replies that it would be absurd to make such a claim. The whole process was simply matching meaningless symbols to equally meaningless rules without any semantic understanding of the contents of the language. Now comes the crux of the argument from the perspective of AI. For Searle, a computer operates solely on syntax. It is nothing but a complicated rulebook. Therefore, if syntax is not enough to explain semantics, and if the computer is a syntactic system, it follows that the computer doesn't experience meaning, and strong AI is impossible (see also Harnad, 1990).

A typical response to Searle has been to say that while the thought experiment may show that the central processor does not understand Chinese, the

whole room, including the Chinese people outside does (Searle, 1980). But this does not reduce the force of the argument which predicated on the question whether syntactic operations suffice for semantics. Identifying meaning in the whole room simply displaces the location of the meaning, but does not provide an explanation as to how meaning comes about, and given that computers function by syntactic operations, there is no way to circumvent the argument by these means. One can't escape from a formal-syntactic system by applying more formal rules (Rosen, 1999), as it this implies an infinite regress (Dreyfus 1972, 2012) or else a positing of a homunculus at the end of interpretation, which will not do as an explanation of the semantics of the system.

Hubert Dreyfus was one of the original (1972) and, along with John Searle, most famous critics of AI. A Heidegger scholar and a phenomenologist, Dreyfus presciently identified certain tacit philosophical assumptions underlying the AI attempt, which according to Dreyfus the AI community was importing without the corresponding historical critique, and thus dooming the attempt to a failure. Dreyfus identified four corollaries stemming from the assumption that man functions like a general-purpose symbol-manipulating device (1972, 68):

1. The biological assumption that the brain processes information in discrete operations by way of a biological equivalent of on/off switches.
2. The psychological assumption that the mind can be viewed as a device operating on bits of information according to formal rules.
3. The epistemological assumption that all knowledge can be formalized such that whatever can be understood can be expressed in terms of logical relations.
4. The ontological assumption that what there is, is a set of facts logically independent of each other.

The core problem, according to Dreyfus, is that the attempt to build a system that acts meaningfully up from a ground of some primitive units, sense-datums, bits, or independent logical facts will eventually run aground if they thus attempt to construct the world of meaning *from* those atomic facts. The problem is that meanings are not in those objects nor in the assembly of those objects. For Dreyfus, *meaning was prior*.

The crucial aspect of the argument for this thesis, is that syntax, rules, computations, or atomic facts do not ground mentality. The ground of mental life is in meaning, however dim in the beginning of organic life (Jonas, 2001), which is subsequently elaborated into forms that acquire syntactic properties. How it is grounded in organic life is by integration that subsequently differentiates, not differentiation that subsequently integrates. Thus, there are reasonable ground for saying that the quest of AI is approaching the problem the wrong way around.

4.3.2 Algorithms

An algorithm is "a precise and unambiguous specification of a sequence of steps that can be carried out mechanically" (Aho & Ullmann, 1994, 5). As Deacon (2013) notes, and as we have implied, the power of algorithms and by extension computers in the inherent agnosticism towards meaning it has. The power of mathematics or logic is in the fact that the syntactic properties of some kinds of reasoning processes are enough to describe it. The contents can be changed and injected into the framework and the algorithm will churn out the correct answer. Of course, a world without semantic contents is quite impoverished, but based on the operating principles of computers, such is the world they inhabit, indeed *must* inhabit (Rosen, 1999). One might argue with some reason that computers do exhibit a kind of semantics when, for example, a neural network is trained to recognize some patterns in datasets which we associate with contents, dogs and cats for example. Such quasi-semanticity can suffice for some tasks, to be sure, but it is not an argument from first principles as we have attempted. Furthermore, the limitations of such systems become apparent not only as technical, but as foundational problems, in the kinds of mistakes made by so-called deep neural networks. Indeed, according to MIT professor Patrick Henry Winston, "no one knows what the neural nets are doing", and "a cottage industry has emerged where researchers try to fool neural nets" (Winston, 2016). See figure 5 on page 30 for an illustrative example.The limitations of computer systems in what might be called meaning or semantics needs to be taken very seriously. We have made the argument here that the problems are not only technical, but go deeper into the foundations.

To take a view from "a believer", and as the reader will recall the person who coined the term artificial intelligence, let us turn to a sober review John McCarthy penned in 2007. According to the paper, the major obstacle between technology at the time and human-level AI could be captured under general capacity to succeed in common sense informatic situations. Such situations are distinguished from *bounded* informatics situations by the impossibility of determining beforehand what counts as relevant fact, and that the facts themselves may be incomplete (McCarthy, 2007). The formalization of situations such that traditional approaches in AI and computer science yield satisfactory results is problematic, given that real-world situations are not often *bound* by rules. Rather, rules in a domain constrain behaviors and afford regularity and predictability, but it is not correct to say that human behavior is *determined* by the rules, as the case would be for the machine. Preliminary studies (forthcoming) conducted on the thinking of experienced ship captains seems to indicate that rule-based behavior is ascribed to others and adhered by the seafarers based on extremely context-sensitive understanding, and the rules are there to provide some structure and tools for decision-making, but hardly fix it. Thus, as opposed to a bounded information situation like chess (Saariluoma, 1992) which can be turned into a computational exercise, it seems possible that autonomous ships determined by ironclad algorithms are liable to be worse than stupid in dynamic common sense

demanding situations. Of course, it may be that it will be the human understanding which will simply incorporate understanding of the behavior of autonomous ships, but this seems to argue for turning the artefact into "not even stupid", given the predictability the would then entail.

The basic problem is this: computer science requires a formal blueprint of what is to be achieved (a function) in order to be implementable at all. This means that the designers abstract out a form of human life of regular action and then cast it down, so to speak, into a causal mechanism. What is left outside is the middle level, the contents and senses from which the abstraction took place. Because there are no actual contents guiding the behavior of the machine, there is an extraordinary pressure and need for fine detail in the other levels. It seems both obvious and unsatisfactory to say that our behavior is guided by the contents of our experience. It is obvious to us, but it seems unsatisfactory from a scientific perspective to take them as givens. As Dennett (1990) noted, the so-called frame problem of how certain patterns acquire relevance for the system is due to the fact that a computer "starts at zero", with no intrinsic understanding of relevance in the world, the achievement of AI rests on the ingenuity and attention to detail from the human programmer. Ultimately this displacement of "real intelligence" to the human creator in this sense makes the systems brittle, but also even at best merely clever software. These kinds of methods cannot yield intrinsic intelligence, and by extension, autonomy.

An understandable rebuttal against our discussion is, as Boden (1989, p. 50) notes, is that an algorithmic picture of intelligence need not imply anything like a simple sequential process. The differences, in this view, are in the various higher-level techniques employed, such as connectionistic architectures as opposed to more traditional AI approaches (such as Newell & Simon, 1961; 1976). But which technique is best for computing a particular function is a different, technical discussion, than the foundational issue we are targeting. Indeed, many algorithms, some more efficient than others, can be devised to implement a particular computation (Marr, 1990), indicating how content and algorithm are somewhat irrelevant for each other. But there is no principled argument, other than by assumption, of how from unambiguous mechanical steps emerge meaning and significance if those are not parasitic on human understanding.

4.4 The Cognitivist Inversion

In order to contextualize the presuppositions outlined before, let us now turn to how the computational-representational picture of intelligence has played out in cognitive science. In the cognitive sciences, mental processes are often defined as computations over mental representations (Frankish & Ramsay, 2012; Thagard, 2005). That is, rather than considering computations as a *sub-set* of mental processes, the cognitivist inversion places computations as the *foundation* of mental processes (Kary & Mahner, 2002; Searle 1990b). This view has deep roots going

back to Turing (1950, see also, Shanker 1998), and beyond to the ideas of Descartes, Hobbes, Leibniz, Frege, and Russell (Dreyfus 1972, 2012). As articulated by Thagard (2005), the conglomeration of these views leads to a complex three-way analogy between the mind, the brain, and the computer. If the mind is to the brain as software is to computer, and if both are at bottom doing computations, a specific case of the thesis of *multiple realizability* (for an overview see Bickle, 2016) follows. That is, it is possible to fully recreate mental processes on computers, and therefore that strong AI (Searle 1980, 1984) is possible in principle by the methods afforded by Turing-machines. The notion that ties these concepts together is that computers and mind-brains are similar in the sense that both are *systems* that *process information*.

But are we justified in saying that mind-brains are at bottom computing? And is "information processing" anything but a "suitcase term", designed to hold and package somewhat disjointed notions together? Searle (1990b) has argued that cognitive science seems to posit a computational-representational level somewhere between the physiological processes of the brain, and the subjective experiences we all seem to have, and that this new level seem arbitrary and confusing. How would we differentiate between the computational processes of the brain as a biological organ on the on hand, and that of say, digestion on the other? Any process can be *described* as a computation, but that does not mean that it *is* a computation. The reason why foundational assumptions behind the attempt to instantiate intelligence in machines were not clearly seen was that they matched, and still do perhaps, so perfectly with the syntactic formalism that underlies modern science (Heinämaa & Tuomi, 1989).

We have sought to make the case that it is unlikely that human intelligence is founded on computations, but that computations are a subset of the capacities of the human mind, and indeed, mind-*dependent*. Thus, it is unfeasible to take such a highly abstract form of human thought, cast it down into a regular mechanisms such as a computer, and expect all facets of the mental, especially the semantic, to miraculously appear.

4.5 Functionalism

According to Revonsuo (2001) the thesis of functionalism forms the hard core of (classical) cognitive science, and according to Bechtel and Mundale (1999) has become orthodoxy in the philosophy of mind. The claim, originating with Hilary Putnam (1960, 1967), sought to stake out a purpose for mental states in between the material processes in the brain on the one hand, and against logical behaviorism on the other, both of which in their own way seek to deny if not the existence, then at least the explanatory power of mental phenomena. The claim of functionalism is essentially this: mental states exist by virtue of their causal roles *among themselves*, and in relation to sensory and motor processes. That is, mental states can cause other mental states, and influence behavior, but are not reducible to

particular configurations in the underlying neurological structures. That is, mental states have *functional* roles in economy of the organism. By introducing the notion of computation, namely that mental states are computational states, a line can be drawn between the physical aspects of an organism and the logical aspects of the program it is running in an exact analogue to a Turing machine, making it possible to describe the system from both the physical and logical laws that govern its' behavior (Shagrir, 2005).

An argument for this, according to Putnam (1960), is in the idea that it seems like mental states, such as pain, seem to be realizable in multiple different types of animals and thus unlikely to be *identical* with the particular nervous systems as such. Thus, if many kinds of nervous systems can realize the same function, it follows that (at least to some degree) mental states are multiply realizable. In other words, the brain states of different species or even different individuals of the same species are unlikely to be the same as they experience pain, and yet we should think that they are all experiencing the same mental state (of pain) whose role is to trigger certain behaviors, such as avoidance of the source of pain in the short- and long-term. The former might be an immediate action, and the latter a more general avoidance of the source. Thus, Putnam argued (and later argued against) that we could be made of "swiss cheese and it wouldn't matter", as long as the functional characteristics would be instantiated (Shagrir, 2005).

By describing mental states as functional, one could say functionalism simply moves behaviorism, the matching of inputs to outputs without regard for intervening mental activities, inside the organism in the guise of programs. Or identifies mentality with those programs. The metaphor of program is perhaps apt to describe certain animals, and of course computers from where it originates, but merely complexifying the program does not obviously or logically conclude with what we want to identify with the mental. It remains on the side of the fence which we call automatic as opposed to autonomous. The cold fact that mental states are intentional and have representational content means, as John Searle (1980, 1984, 1990) has argued, that mere functional specifications by syntactic rules do not account for the arguably key feature of the mental. The abstraction entailed in such a formal specification has its' uses, but its' power is its' handicap: it leaves the mental contents behind (Saariluoma & Rauterberg, 2015).

This is not to say that mental states shouldn't be characterized by their functional roles, but that this does not *exhaust* their nature. It is likely true to say that mental states are characterized by the relations among each other and connections to sensory inputs and motor outputs. But there is something more, and that something more is absolutely central and it is intentionality, meaning, and mental contents. As Shagrir (2005) points out, the formal syntactic operations of a program run over a finite set of symbols, but symbols attain their power in computational terms from their arbitrariness: the extension of the symbol "1" is the number one, but it is by way of convention and our ascribing it the content "number one" that it has the extension it does. Formal programs that do not ascribe contents to symbols lack semantics, and therefore meaning, intentionality, sense,

and so forth and thus can't be properly conceived of as mental (Searle 1980, 1984, 1990).

Our discussion is constantly running up to the necessity of understanding the role meaning and contents play in our mental life and consequently the importance of instantiating similar abilities in machines for AI to reach its' goal. It is also running towards a trap that has long been considered a fallacy: the homunculus, or little man in the head (Deacon, 2013). For now, let us say that it is only a fallacy if posited as an unexplained explainer to prop up some other theory. Here we are indeed running directly towards it, but with the attitude that it is precisely what needs to be explained, rather than explained away or ignored, because as the history of psychology and cognitive science have shown, it will not go away by ignoring it. As the history of AI has shown, our inability to conceptualize how to even begin conceptualizing how matter in motion leads to meaning and significance (Dennett 1986; Fodor 1985) is limiting practical success, insofar as those are tied to common sense (McCarthy, 2007) and therefore proper intelligence and autonomy.

4.6 Computations are Multiply Realizable, but are Mental Processes?

Given the vast variety of different systems that can be used as computational devices, it seems clear that computation is multiply realizable. A Turing machine can be constructed out of wood (Ridel, 2015). A person can count "in their heads" or by using their fingers or the abacus or on a piece of paper. The only thing that is required is that the system in question can assume discrete states in an orderly fashion. Your fingers can be of many lengths and sizes, but raising two fingers can for an observer be sufficiently well discriminated to reach the conclusion "two". Of course, and this is more than humorous, the observer must understand that the fingers are raised in the context of calculation, lest the gesture be interpreted as signifying something else. Clearly, the human mind can assume discrete enough states to implement computations. But there is a clear difference 'between being able to assume' and 'being constituted by'. It is plausible, as Fodor (1985) suggested, that the limits of "Turing style rationality" are rather narrow indeed, despite the power they entail. In other words, it is likely that while computations are multiply realizable, mental processes as a whole may not. Or put another way, it may be that computations are only a subset within a far more expansive landscape, and computations, in minds, fingers, or machines, only make sense within that vast landscape. There may be orderly change, and orderly change has some similarity with computations or what might be called physical laws, but for them to mean anything, they have to occur within a framework of human understanding. Thus for properly informational processes, the physical substrate that can assume certain regularity is necessary, but not sufficient.

In fact, it may be more plausible to say that brains, much like computers, are being hijacked by properly informational, i.e. mental, processes. This is not to say that they descend from outside, but that the mental emerges somehow from concrete physical system typified by humans, and once it emerges it can, occasionally, hijack the physical substrate from which it emerges, and by extension many others in the external world.

4.7 Summary

The purpose of this section was to bring the endeavor called artificial intelligence (AI) into focus as the most natural waypoint after connecting autonomy and intelligence in the first section. We have tried to make clear the most basic operating principles of the computer as it relates to the question whether it appears as a suitable platform for the instantiation of intelligence. The answer has been mixed, in that it seems clear, already from current machines, that some tasks that previously required human intelligence can indeed be mechanized and implemented on a computer. It does not, however, follow that all tasks fall into this category. It does seem, that AI struggles with precisely the aspects that one would expect it would, given the operating principles of computers: the mechanization of abstractions. Thus, we feel justified in exploring whether AI, as predicated on the computer, can be a field that studies the principles that make intelligent behavior possible in natural or artificial systems (Poole & Mackworth, 2010). This is not a wholesale dismissal, but an intuition to examine more closely the difference between computer models of intelligence and the "real thing", judging it not from a prescriptivist point of view, but from deeper operating principles.

The story of artificial intelligence has been one of booms and busts (Dreyfus, 2012). The community has a tendency to get excited over various techniques from rule-based systems to neural nets, and back again. Such techniques are of course meaningless without contents over which the techniques are to operate (Chandrasekaran, Josephson, & Benjamins, 1999). However, a formal ontology specifying a taxonomy of categories and items within them is but an abstraction without a principled explanation of how the contents fill the system (other than by human operators).

Our conclusion has been that there is no system-internal methods in computers by which those contents can arise, precisely due to the operating principles of computers: the mechanization of abstractions. Furthermore, it is thus no surprise that where AI systems most struggle with are with situations where the appreciation of contents is required. A formal system can be approximated by machine techniques given sufficient computational power. Informal, so called common-sense informatic situations (McCarthy, 2007) are far more difficult, because due to the unavailability of actual semantic contents, an extraordinary pressure is placed on fine adjustment on the formal side of things. Even when successful, it can be argued that it still remains essentially brittle and thus fundamentally unreliable. But there is further significance. We brought out how even

theories of the mental have trafficked with some of the same presuppositions as AI. If computations and functional explanations of mental contents are to be demoted as only facets of the mental, it follows that theories built on those foundations must go as mostly descriptive and essentially pseudo-causal (Saariluoma, 1997).

5 MULTIPLE REALIZABILITY

In cognitive science and AI, multiple realizability refers to the contention that a mental kind can be realized by many distinct physical kinds (Bickle 2016). For example, an extreme version, such as what John Searle (1980, 1990) has referred to as "strong AI", contends that one could in principle fully recreate mental phenomena on a computer. A less bold version might say that *some* mental processes are such that they are at least on a superficial level capable of being replicated on a computer. An obvious example might be the calculation processes in a pocket calculator. At least judging by the result, we could say that given a task of calculation, the pocket calculator by far exceeds human speed and accuracy. Indeed, understood in highly abstract way, from the perspective of the *task*, one could say that the same capacities are being realized in the machine and in man. The crucial task, especially in the context of autonomous technical artefacts, is to describe the extent to which, and what kinds of, mental processes that currently go on in vivo can be reproduced in silico (Bunge, 2004). In other words, what are limits of multiple realizability.

Two primary questions relating to multiple realizability are 1) what (our conceptualization of) is being realized and 2) how (what are the laws and immanent forms of determination by which) it is realized. To follow Bunge (1979), it is important to distinguish between laws and law statements: the former being the actual patterns of being themselves, and the latter being our conceptual and epistemological attempts to grasp those patterns in the forms of scientific laws, for example. The important question is whether it is necessary for the instantiation of intelligence in artefacts to grasp its' lawful antecedents in natural systems: is it necessary to discover the (antecedent) laws of the mental before they can be invented? Given the counter-intuitiveness of scientific discoveries, it is hardly to be expected that the so far occult principles by which mentality emerges with life, or life itself for that matter (Rosen, 1999), will obey the intuitions with which we make our ordinary way through daily life such as simple causality or intentional language. Insofar as the mental is both determined by some antecedent conditions and levels or scales but also capable of to some extent being a cause unto itself and directing the behavior of the organism, we come upon tricky territory, but one that lines quite closely with the questions of autonomy and automaticity. That is, insofar as we discover or invent an autonomous system it creates along with it ipso facto an autonomous science. This has been the case even with computer science, that is, it seems we can discover certain things about computers and their capacities and limitations even though they are wholly artificial phenomena.

Furthermore, as design and science has important epistemological characteristics the conceptualizations, characterizations, and representations they employ make all the difference. Changes in representations and representational languages can make certain problems less vexing; the ways in which we conceptualize problems afford different approaches to solving them (Simon, 1981). All

concepts have some limits in their scope of expression (Saariluoma, 1997; 1999). Accordingly, some problems may not be hard as a consequence of our inability to solve them, but our inability to articulate them in a manner that makes them solvable, or less vexing, or disappear altogether. It is at least plausible that our inability to account for key features of the mental such as subjectivity, semantics, and meaning reflects necessary consequences from our assumptions and representational languages and concepts. And furthermore, this may be connected with how we struggle to instantiate such features in artifacts.

5.1 Pluralism before Emergence or Reduction

Philosopher Karl Popper (1979), perhaps best known for his examinations of science, held the view that any ad hoc reduction by linguistic means of a plurality to some monism is dangerous for the advancement of knowledge. Quoting Imre Lakatos, he warned against 'degenerative problem shifts', which eliminate possibilities on making advancement on genuine problems by eliminating them by fiat. Case in point is the reduction of human thinking to the processes of the brain by a wrong-headed application of Occam's Razor by positing only material processes to account for psychic processes. A reduction between different fields, most notably chemistry and physics, is sometimes possible, but it should not be assumed. A naturalism, for Popper, should proceed from a pluralism.

The kind of reduction most applicable to AI is one of reducing the mental to computations, and posit thus their multi-realizability in computers and in man. We have seen however, that this position seems doubtful.

The mental or the psychic occupies a distinct level of reality, signified by the different sciences that grapple with phenomena that appear at that level; by the way we ordinarily talk and explain human behavior; and of course by our own subjective experience. The fundamental tension is in how we should conceptualize this level of reality in a manner continuous with a broadly naturalistic perspective while retaining and explaining its' curious characteristic.

5.2 Pluralism and Modularity

Taking a step forward after acknowledging pluralism in the sense here that the mental is one real category among the ontology of reality, we may consider plurality within the domain itself. This is the question of modularity versus holistic or central (information-processing). The idea of modularity was perhaps most famously introduced to the discourse of cognitive science by Jerry Fodor (1985). For Fodor, a mental module had certain key characteristics: Domain specificity; Mandatory operation; Limited central accessibility; Fast processing; Informational encapsulation; 'Shallow' outputs; Fixed neural architecture; Characteristic and specific breakdown patterns; Characteristic ontogenetic pace and sequencing.

Indeed, such an account lines up rather well with what computers do, and indeed how some "modules" in our mental ecosystem function. When combined with the idea of (near-)decomposability of a physical system, introduced by Simon (1981), the idea follows that mentality (as modular) can be identified in a similarly modular brain, composed of complicated but ultimately simple smaller specialized information-processing mechanisms. It further follows, that perhaps equally such simple processing units could be instantiated in a computer, yielding intelligence in the machine. The fundamental presupposition, as identified by Schierwagen (2012) is thus one of decomposition and localization: particular, say, capacities of a system can be localized, and the system can be decomposed roughly along those lines and the (relatively) weak interconnections among the modules. But as Fodor (1985) acknowledges, these kinds of mental processes are but one part of the whole story. What he calls "global" integrated processes that can integrate various streams of information and assign meaning and content to them are arguably more important as far as special qualities of human minds are concerned. That is not to say that the "global" processes would not depend in some sense on the simpler modular processes, any less than all of these depend on more basic metabolic processes that sustain life, but that while explanations targeted at these simple modules are useful and interesting, they should not be allowed to take over the entire explanatory framework. The mind may be modular, in the sense that it is differentiated both in structure and in experience, but the far more important and neglected facet is its integration. Even "organic logic" attests to this: life differentiates from a simple unity both per individual organism and over evolutionary history. Thus, by naturalistic assumption, the same applies for mentality.

5.3 Supervenience and Emergence

Supervenience is a conceptual tool for grappling with issues that emerge from the apparent multi-leveledness of reality, most specifically to grapple with the problems that emanate from the distinction between mind and matter and their relationship. It has a family resemblance to the concept of emergence in that both attempt to deal with levels of reality, but typically supervenience seems to hold that there is a strict dependence (a supervenience) of the lower levels to the higher levels such that changes in the lower level always mean changes in the upper level (and vice versa), but such that the arrows of causality flow upward from the physical base (and within it), rather than downwards from the upper levels (Jaworski, 2016). With emergence, the focus is more on how the emergent properties, here the mind, can have causal efficacy in the world. We could say that higher-order phenomena emerge from lower levels and their relations are supervenient on one another either in the strong or weak sense of the term. The term has a checkered history, but its' introduction to contemporary philosophy can be credited to Donald Davidson (2005) (see also McLaughlin & Bennett, 2018):

> [M]ental characteristics are in some sense dependent, or supervenient, on physical characteristics. Such supervenience might be taken to mean that there cannot be two events alike in all physical respects but differing in some mental respect, or that an object cannot alter in some mental respect without altering in some physical respect. (Davidson 2005, 116).

The relevance of the notion of supervenience for our thesis should be clear. It seems like strong AI (the idea that it is possible to recreate minds on computers) must also take a very qualified stance towards supervenience. It must hold that, yes, mental properties (understood as information-processing) supervene on the physical substrate but only in a relatively weak sense given the obvious distance between computers and human beings as physical systems. This leads directly into the problem that since almost anything, such as strings and coke cans, or wooden contraptions (Ridel, 2015), can be assumed for the function of computation, it would follow that any physical substrate will do for the mental. But as our discussion illustrated in the previous section, such a view is implausible and issues from the mistaken identification of the ground of mentality with computations, whereas we argued for the reverse.

In the context of information, the problem of supervenience becomes clear. We can take it to be the case that there is some necessary connection between the physical substrate on which the information is instantiated: the words on the screen have multiple layers of physical processes which supervene on each other, if the computer malfunctions or the battery runs dead, the physical substrate on which the letters are founded will momentarily cease to exist. On the other hand, that the words *mean* something, that they can inform the reader or articulate a thought is dependent on an observer. Thus, in order to be something more than physical marks on a screen or a piece of paper, they entail a particular kind of relationship to an observer which makes them information. Is this relationship one of supervenience? It is in the sense that the specific information they are intended to convey or articulate will not be available if the letters disappear. But there is no necessary connection with the information and the letters and sentences, indeed, especially in the context of this thesis, grappling as it is with highly abstract and complicated issues, it is quite likely that the reader will interpret different information from the text, based on his or her knowledge and dispositions. There are more patterns in the implications available than either the reader or the writer is likely to identify. Thus, there is no strict one-to-one correspondence between the physical medium and the information it can convey. Furthermore, we could say that the information could be within certain limits be conveyed by quite different sorts of media: they could be printed with ink, read on a digital screen, written by hand on a piece of paper, conveyed in spoken language, or written in the sky with a stunt plane without changing the essential properties of the message. Does this then mean that similarly mental properties are multiply realizable? Only if one could show that the crucial part of the equation, the interpretation of the information, which makes it information, is itself multiply realizable. It is precisely here that the concept of information with its'

possibility of multiple interpretations as part of its' nature problematizes supervenience, or any simple and strong correspondence between the mental and the physical. Furthermore, it is in this sense that indeed information and information-processing is multiply realizable, but dependent on an observer. The question is thus whether the observer as such is multiply realizable, and in what sense? This is a most difficult question, but it seems from the preceding discussion to be a vital component. As a reader familiar with these lines of thought may recognize, what we are here identifying with information is traditionally identified in the discourse with intentionality, the way in which mind seems to be projected towards, or *about* things (Brentano, 2009; Jacob, 2014). The reason we are conflating or merging the terms is that it brings out a certain irreducibility between intention and information which furthermore shows that the problems facing AI and cognitive science are not so distant from each other, and the task may be more than devising ever more complex programs or clever algorithms to model mental phenomena.

Emergence is a concept that describes when interactions at a lower level give rise to phenomena that obeys its' own laws (Bechtel, 1994). Here the important question is whether those emergent properties can be considered real entities, or simply convenient shorthands – reducible to the interactions of elements at a lower level and without any intrinsic causal capacities. In the context of cognitive science, if the mental is an emergent phenomenon that comes about from the interactions within the brain in interaction with the environment, does it have a status as a semi-autonomous entity or is it, in principle, completely reducible to brain states without residue?

One may approach the question with an analogy to temperature. Temperature is a measurable property of a whole, which is the result of mean kinetic energy at the level of molecules that compose the whole. It is thus a property that emerges from well-known interactions at a lower level, but retains a certain autonomy given that it can be attributed as a *property* of a system (Bunge, 1977). Indeed, there *is no* temperature apart from its' macroscopic manifestation, even though its' emergence is explainable by microscopic interactions.

This is important in relation to the autonomy of the special sciences. It means that there may be an explanatory story to be told how macroscopic properties emerge from lower level interactions, and yet the macroscopic level may have an autonomous existence, indeed the being of the phenomenon is to be found only on that level, irrespective of any explanations of *how it came about*.

Now, one may ask if this would imply that, for example, intelligence and mental life could be explained by microscopic interactions among some logical units, neurons for example, that *add up* to make "more" or "higher forms" of the same basic operations? The most simple way this does not "add up" is that while in some regards more neurons (or more interactions) do mean more intelligence, it is clearly not the whole story even if we would attribute intelligence only to the brain, simply because far more important that numbers seems to be structure and dynamics, as far as brain function is concerned. Far more crucial for intelligence

seems, even from the perspective of brain science, seems to be structure and form which directs or captures certain forms of determination.

The essential problem of supervenience and emergence was put succinctly by Jaworski (2016) in the context of action:

1. Actions have mental causes.
2. Actions have physical causes.
3. The mental and physical causes of actions are distinct.
4. If action have multiple causes, then they are overdetermined.
5. Actions are not overdetermined.

The essential point is that if one posits a distinct mental cause for an event, say my hand going up, and a distinct physical cause for the same event, it seems like one is overdetermining the event. This follows from an overly simplistic notion of causality, of which more presently, but the way in which Jaworski (2016) seeks to resolve this tension is interesting from the perspective of this thesis. He seeks to conceptualize the mental causes of an event as structures that constrain causal events closer to the immediate behavior, say, of raising an arm. Thus, if the causal efficacy of the mental is seen not as of a different kind of substance somehow causing physical events, but through the lens of orders or levels of complexity which constrain, rather than directly cause, physical events to occur in a certain manner, we may have a method of escaping from the problem. Constraints offer a kind of gestalt switch from doing as in causing every detail, towards orders of constraints on possible behavior. How information, which we understand less well than often thought (Checkland, 1994), can constrain behavior or have semantic properties is an open question (Deacon, 2013).

We will next turn to a discussion on different forms of determination, one of them mental. The crucial terms are indeed *form* and determination.

5.4 Forms of Determination

Recall that autonomy comes from the greek term autos (self) and nomos (law) and means *control of the self* (Bateson, 2002), or more literally *having its' own laws*. Its' antonym is heteronomy, which means other-governed (Thompson, 2007). Metzinger (2017) defines autonomy as the "capacity for rational self-control of overt behavior", while using the term *M-Autonomy* to refer to the capacity to apply self-control over one's *mental* functions, a kind of second-order autonomy. In a similar vein, for Margaret Boden (2008), an individual's autonomy is the greater the more it controls its own behavior via self-generated inner mechanisms that are available for reflection and modification by the individual itself. Thus, an autonomous system is both responsive to particular problem-situations and reflexive in terms of the mechanisms that control the responses and in terms of the wider context (Boden, 2008). In the context of maritime vessels, Insaurralde and

Lane (2013) define autonomy as "the ability of a system to govern itself by making decisions, implementing the choice made, and checking the evolution of such actions taken."

In biological systems, one can plausibly make the case that cognition has emerged in evolutionary history as an adaptation predicated on the *biological autonomy* of living systems that need to maintain a precarious existence via metabolical exchange with the environment (Jonas, 2001). Thus, cognition increases the range of possibilities and autonomy of the organism in order to extend its' possibility of maintaining biological, lower level, autonomy (Jonas, 2001). The point here is not to argue for a necessary connection between lower and higher-level autonomies, but simply to note that, certainly in humans, cognition and autonomy are overlapping and complementary systems (Vernon 2014). Equally judging by Insaurralde and Lane's (2013) definition (see also Williams, 2015), insofar as autonomy entails self-control and decision-making, the same applies for ships. In short, in the context of ships autonomously sailing the oceans and straits of the world, no autonomy without intelligence.

In this section, we wish to address the question from a different perspective and attempt to connect intelligent self-control with forms of determination operating with increasing orders of constraints.

5.4.1 Forms of Determination

> We ought then to regard the present state of the universe as the effect of its anterior state and as the cause of the one which is to follow. Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective situation of the beings who compose it – an intelligence sufficiently vast to submit these data to analysis – it would embrace in the same formula the movements of the greatest bodies of the universe and those of the lightest atom; for it, nothing would be uncertain and the future, as the past, would be present to its eyes.
>
> — Pierre Simon Laplace, (1902, p. 4)

Laplace's example succinctly captures what might be called the doctrine of causal determinism (Bunge, 1979). What makes mental causation a *problem,* is arguably the view that the universe contains only one form of causation, physical, and into such a closed, gapless, world of causal chains no mental (or arguably informational) forms of determination can intervene (Burge, 2007). This leads to the position of dismissing the mental as a distinct ontological kind or perhaps of positioning it as epiphenomenal without any causal powers to intervene with affairs in the world, even the body itself. We shall not argue for the attribution of any soul-like properties to the mental, but neither shall we dismiss its' causal efficacy in the world. This efficacy is limited, and acts by proxy with the environment or by quasi-logical connections within itself, but its' nature is "no more" mysterious than that of information. Neither information nor the mental have only physical properties, but few (cognitive or computer scientist at least) would argue against

the proposition that information can't play a causal role in the world. The question is not when and where mind emerges (distinguished for example by consciousness and mental representations), but when does information? That some deep, even fundamental connection exists between information and the mental has been known for a while (Wiener, 1985), but one may yet question if we have a thorough grasp on what information *is?* What marks informational kinds apart from, say, energy and matter, and what are the connections between them? Furthermore, even the example by Laplace is essentially about a knowing subject, and one would surely be tempted to say that if such a being were to exist and have all the information at its' disposal, it would ipso facto be able to intervene on some causal chain of events and thereby altering the course of the vast clockwork. Thus, Laplace's universe is valid only in a universe without a knowing subject, rendering the whole thought experiment paradoxical.

A crucial concept that is buried in the notion of forms of determination (Bunge, 1979) is that of *form*. Namely, no one within a broadly naturalistic worldview is looking to deny the penetrating presence of natural laws in some pockets or regions of space. But within such a framework, one can still move towards understanding the forms and structures within which causal, even wholly deterministic processes can be channeled. Isn't this much the whole ethos of engineering design to begin with: the organization of matter, energy, and signals to perform some function (Pahl, Beitz, Feldhusen, & Grote, 2007)? In broad terms, it is the imposition of *constraints* on the flows and patterns of matter-energy-information.

This is the key concept. In artificial systems, these constraints are imposed from the outside by the human mind and hands. In the subset of natural system that we call living, those constraints are imposed and maintained by the operating principles of the *system itself*. In the advanced organisms, especially humans, there is a meta-level of constraint-imposition that is achieved by what we can generally call information and processing thereof: mental and cognitive forms of determination.

5.4.1.1 Autonomy as a Form of Determination

Determination and autonomy form a unity. Autonomy, broadly defined as an agent's capacity for rational self-control of behavior (Metzinger 2017), can't be instantiated in a universe without reliability, and therefore determination within constraints, which forms a prerequisite for a system or mechanism to do work (Deacon, 2013). From this perspective, pitting determinism and autonomy (a form of "freedom") against each other is a false dichotomy. Rather, determinism and autonomy are different levels on a dependent hierarchy (Wilden, 1987) with the latter emerging from the former. This idea is echoed in Bunge (1979, 20) as well: various forms of determination form a hierarchy of increasing complexity where all but the first two (quantitative self-determination and causation) are grounded in the lower types. Here we wish to draw attention to the need to make distinction between autonomy and causal determinism, using biological autonomy as the exemplar (arguably the only exemplar in the natural world). Bunge

(1979) sought to distinguish among *different forms* of determination, essentially arguing against the idea that all phenomena could be collapsed under the rubric of *causal determination*. For Bunge, causal determination was but one (he identifies at least eight) form of determination on a hierarchy where the higher types, such as statistical or teleological, *depend* on the lower types, but are not entirely *reducible* to them.

Table 4: (some) Forms of Determination according to Bunge (1979)

Form of Determination	**Definition and illustrations**
Quantitative self-determination	The determination of the consequent by the antecedent. Illustrations: a) the successive positions of a freely moving macroscopic body are uniquely determined by its position and velocity at any prescribed instant of time.
Causal determination	Determination of the effect by the efficient (external) cause. Illustrations a) If a bullet is fired against a window, the glass is broken. b) If an electro-motive force is applied to the ends of a piece of metal, an electric current is set up in the metal in accordance with Ohm's law.
Interaction	(or reciprocal, or functional interdependence): determination of the consequent by mutual action. Illustrations: a) The orbits of components of a double star are determined by their gravitational interaction. B) The functioning of every gland in the human body depends on that of the remaining glands.
Mechanical determination	of the consequent by the antecedent, usually with the addition of efficient causes and mutual actions. Illustrations: a) Forces modify the state of motion of bodies (but motion may exist before the application of the forces). B) The streamlines in a fluid are determined by the latter's previous state, by the external forces acting upon it, by internal friction (viscosity), and by internal pressure differences.
Statistical determination	Of the end result by the joint action of independent or quasi-independent entities. Illustrations: a) In the game of dice, the long-run frequency of the event "throwing two aces in succession" is 1:36 b) about one-half of newborn children are females. As in the case of other categories of determination, statistical determinacy may emerge from processes on deeper levels, in which still other categories of determination are involved.
Structural (or wholistic) determination	Of the parts by the whole. Illustrations: a) the behavior of an individual (a molecule in a fluid, a person in a social group) is determined by the over-all structure of the collection to which it belongs. B) the functioning of an organ is partially determined by the needs of the whole organism. But, of course, the whole, far from being prior to its members, is in turn determined by them.
Teleological determination	Of the means by the ends, or goals. Illustrations: a) Birds build their nests "in order to" safeguard their young. B) Standardization is adopted in industry in order to lower production costs. Needless to say, goal-directed structures, functions, and behaviors need not be purposefully planned by anybody.

(continued)

Table 5: (some) Forms of Determination according to Bunge (1979) (continued).

Dialectical determination (or qualitative self-determination)	Of the whole process by the inner "strife" and eventual subsequent synthesis of its essential opposite components. Illustrations: a) Changes of state in matter in bulk are produced by the interplay and final predominance of one of the two opposite trends: thermal agitation and molecular attraction. B) the contrasting economic interests of social groups determine changes in the very social structure of such groups. In opposition to quantitative self-determination, internal dialectics involves qualitative changes. And, needless to say, it has nothing to do with logical contradiction.

The similarity between computations and the operating principles of machines and computers and the doctrine of causal determinism should not evade us. This is again pointing directly at what we have identified as the difference between automatic and autonomous. An automatic process is from a conceptual perspective a causal-deterministic process. But as Bunge (1979) noted, that is but one form of determination in a hierarchy of types. All of them have a reality, and it may be that what is constraining progress in autonomous technology, artificial intelligence, and, conceptually, cognitive science is the way in which we think and conceptualize the issues.

Note how Bunge's (1979) hierarchy implies how the higher types are grounded in the lower types. It is important to note that being grounded does not mean that it can be reduced (to the lower types). There is no need to consider monism as equivalent to clarity - on the contrary. Davidson's anomalous monism is a position that allows for a monistic view of being, but also plurality in its embodiments (Davidson, 2005). We simply need to be clear of the properties (of say, an agent) we are examining and apply the appropriate tools for that level. As Saariluoma (1997) noted, just because causal explanations are proper for almost all questions asked in science, we can't deduce from that that it is appropriate everywhere, and the problem is that since there are no assumption-free theories, when we take causalism as a given we tend towards asking questions that have causal explanations, which may limit our understanding. Indeed, to quote Kenneth Burke, it may be that "much what we take to be observation is simply the spinning out of possibilities implicit in our particular choice of terms" (quoted in Anton, 2011). As mentioned previously, causal determinism seems inappropriate to deal with mental contents (Saariluoma, 1997), intentionality, and therefore agency. Agency is a form of goal-directed behavior, which can be subsumed under the general category of teleonomy. Next, we must turn to some distinctions between forms of teleonomy, itself a part of the category of forms of determination.

5.4.2 Teleonomy

The Greek word *telos* means end or goal, and teleological means end-directed (Mayr, 1974). The word teleonomy was adopted by Ernst Mayr (1974) in his attempt to establish a middle-ground between mere mechanism and purpose in biology, implied by the word teleology (Deacon, 2013). Merriam-Webster defines teleonomy as "the quality of *apparent* purposefulness of structure or function in living organisms due to evolutionary adaptation" (emphasis added). Thus, the term was designed to account for seeming purposefulness in life and evolution without recourse to any grand designer, or the assignment of actual *telos* to the processes of nature which teleology implied. The greek combining word *–nomy* implies mere lawlike behavior and thus it could be used to describe behavior that were oriented towards a particular target state, even without any explicit goal-representation (Deacon, 2013). Von Wright (2004) used the term quasi-teleological to address the exact same issue, to capture and distinguish teleonomic from teleological explanations. Notice here, again, the seeming implication that the orientation towards a particular state of the thermostat and corresponding behavior is now easily treated as similar in kind to a person learning to play the Toccata and Fugue in D minor (Shanker 1998, 50). But Mayr (1974) was more subtle than that. For him, a new distinction was needed. He distinguished between processes which reach an end-state caused by natural laws as *teleomatic*, and processes whose goal-directedness was controlled by a program as *teleonomic*. The problem with this explanatory schema is again the (rather explicit) causalism entailed in the programs by which teleonomic processes are deemed to operate. It reduces the question of human agency and mental contents to an explication of the programs that run in our nervous systems, with the position of the programmer now relegated to natural selection. But isn't it the case that to some limited, but crucially important, extent human agency is characterized by the capacity to "program itself"? Hasn't the person who has mastered the Toccata and Fugue in D Minor "programmed" him or herself in some sense (Shanker 1998)? The natural reply which comes to mind is that while that may be true, it is also just a program among others, one capable of programming other parts by controlling behavior: its programs all the way up. This of course explains nothing, just states an explanatory schema: find the programs in the nervous system that can constrain and direct an organisms behavior towards some goal-states. But such an explanation has the danger of being woefully unsatisfactory, and somewhat descriptive, rather than explanatory – and worse than that, implies an infinite regress if programs are interpreted as rules (Dreyfus, 1972). The basic problem is that while a program is as a metaphor suitable as an explanation for certain kinds of behavior, it does not follow that it is a suitable metaphor for all mental phenomena.

What we should now have in view is a rough description that seems to support the idea that reality consists of various forms of determination, which can be arranged in a hierarchy of sorts, whereby the lower types of determination ground the higher types. The higher types we have now identified within the

broad category of teleonomy: the lawful behavior towards some end-state. What is crucial now is how we can distinguish between the lowest type of teleonomy, mere mechanical tendency towards some state, and highest type of teleology, which we have identified with autonomy and agency – a phenomenon whose instantiation in machines is the ultimate target of this thesis. While keeping in mind that we still consider each form of determination to arise from the types below and thus the relationship between the two must also be addressed.

Consider the classic example of the thermostat. It's behavior is teleonomic insofar as it "aims" to keep the temperature steady by a process of feedback, in which if the temperature of the room dips below some point, a bimetallic strip's reaction to the temperature is exploited in such a manner that it closes the circuit which turns the heat on, which causes a raise in temperature that eventually opens the circuit again, and the cycle is ready to repeat. The behavior of the system is fully causally manifest, and explainable in language of cybernetics or systems-thinking (Churchland, 1994). Here, of course, an attentive reader will note that the reason a thermostat exists in the first place is due to human ingenuity and intention and thus its' regular law-like behavior is hardly anything but the displaced intentions of the designer (and the user). Consider now an example from the animal kingdom which nicely illustrates the differences between human intentions and capacities from seemingly complex, but ultimately dumb behavior. Dennett (1990) describes the behavior of the Sphex wasp, whose reproduction strategy includes digging a burrow for eggs, finding a cricket and paralyzing it with a sting, bringing the cricket to the burrow, checking the burrow, dragging the cricket in and laying the eggs, never to return. The cricket acts as fresh food for the hatchlings as they emerge. When the behavior was studied, it was noted that if the paralyzed cricket is moved between it having been brought to the burrow's edge and the wasp going in to check the burrow, the wasp would seem to get caught in a loop. It would bring back the cricket to the edge and go in to check the burrow. This process would loop as long as the cricket was moved after the wasp entered the burrow. In this instance, it would seem appropriate to use the term program to describe the wasp's behavior, a goal-directed, rather complex behavior that nonetheless can with mild but crucial environmental modification get caught in an endless loop. But consider now the human experimenter who has deemed it interesting to manipulate and probe the wasp's mental life. Even if we extend the metaphor of the program to insects, we certainly see here a breakdown of the same explanatory method. Again, we notice that what marks the difference between human teleological behavior and the teleomatic or teleonomic processes of nature is some kind of access to and awareness of our "mental programs", and the capacity to "self-program" to some extent - if we wish to use the term. It is the capacity impose higher-order constraints on behavior, and also to learn, sometimes highly arbitrary new behavior patterns. Following von Wright (2004) we may now distinguish, crudely, between the *actions* of man and the *behavior* of (perhaps not all) animals (and even physical processes or man-made tools).

5.4.3 Action

It is often said that human behavior or that of living organisms in general is fundamentally different from the processes of nature as conceived in the natural sciences such as physics (Taylor, 1965). We say that a non-living system behaves in accordance to some causal or statistical laws, but that living organisms, especially humans and higher mammals, act in a goal-seeking manner. We say that the reason for a certain behavior is a goal that may have various causes (such as dehydration as a cause for thirst), but that thirst as such does not specify how it is to be satisfied. Sometimes a distinction is drawn (Dretske ,1988, p. 5) between *behavior* and *action* to delineate those behaviors that do not require or involve explicit goals and intentions from actions that do. For instance, the growth of hair or the pumping of the heart would be considered behavior, similarly perhaps even reflexes. But walking to the music store to buy new strings for guitar, replacing them, and practicing "Smoke on the Water" is a series of *actions* that the agent does voluntarily, deliberately, and intentionally. In other words (Metzinger 2017), actions can be distinguished from behaviors by having conscious goal-representation play a central causal role which we apprehend subjectively through qualities such as the sense of agency, effort, goal-directedness, self-control, and ownership. Furthermore, actions carry with them conditions of satisfaction (Searle 1980). In other words, an agent may fail to achieve the goal implied by the action. Actions often have therefore a wider temporal and spatial "reach" as they string together behaviors (characterized usually by automaticity and decreased context-sensitivity) which act as the building blocks of a process that is deemed to approximate the reaching of a goal. Von Wright (2004, 86) terms behavior which has a genuine teleological explanation as action-like. For von Wright, such an action normally presents two aspects: and "inner" and an "outer". The inner aspect corresponds to the intention or will behind the action. The outer part can be divided into two phases, immediate and remote. The immediate outer part of actions are for example the muscular movements that are the proximate cause of, say, a window opening, which is the remote phase of the action. It is interesting to note, that the intentions usually correspond to the remote part of the action, and indeed the bodily muscular movements that are necessary for the action are often quite transparent to us. In skilled action, our awareness, so to speak, transcends the bodily aspects necessary.

But notice here that the intentional arc which we assume directs our behaviors does not seem to be connected within itself by causal laws, but by something like logical, functional, and senseful attributes. Thus, as noted by Saariluoma (1997) the contents of mental representations have senseful or functional explanations. The overt behaviors of which the process of opening a window consists may indeed have a corresponding causal train within the nervous system, but notice again that the way we would explain the process of opening a window is usually in terms of some senseful logical connection, for example to cool down the room or refresh the air. As such the reasons for our behavior are in these

illustrative cases not hidden from us within some occult neural program. People managed to get about just fine before any real understanding of the underlying neural processes were scientifically explicated, precisely because behaviors are not strung together into actions by causal, but by logical and functional connections which have causal connections only by way of us thinking they do. Thus, the relationships between pieces of behavior can be said to have a logical connection, but this connection is not exactly causal, but abstract and logical. In other words, the intentional arc that corresponds (somehow) with overt behavior is informational, not (only) physico-electro-chemical. For only through the term information does it make sense to say that "things" have logical, senseful, or functional connections. This is not a question of which, but how the two connect. If we accept this step in the investigation, we must now turn to a better understanding of the concept of information.

5.4.4 Information

Information is in colloquial use a rather nebulous term. In this thesis, we will follow Norbert Wiener's (1985) original intuition and seek to recognize it as a third domain, in addition to matter and energy. See also Popper (1979).

Information, from this perspective, is not simply order and pattern in physical medium, nor is it something wholly ineffable. It occupies, as noted by Gibson (1966), a curious position in-between. The ecological psychology debate over how much the information is in the medium that conveys it is neither here or there for this thesis, although very interesting. What is important here is to recognize that information is not something intrinsic to the signal medium, nor is it completely removed from it. Information occurs in a relationship, but between what? If we look to the signal medium for anchorage, we run into trouble given that coded variety, such as a written language, is only a typical form of information. For a captain on a ship the *absence* of something can be informational. The radio signal not received, or the sound not heard from the engines can both be a reason for attention and action for the experienced seafarer. And even noting that information exists in a relationship, it yet seems to retain a curious autonomous status. As Popper (1979) noted there is a sense in which information has its' own laws. Information, unlike energy-matter, can be destroyed and created (Anton, 2012). Information, properly conceptualized, marks the difference between mental determination and causal determinism. But information is like the word animal, without some taxonomical categorization the term is too broad for specific discussion.

To get a grip on the diversity of facets on the notion of information, consider Winning and Bechtel's (2016) identification of five dimensions along which interpretations of information may vary: type of intensional content; type of vehicle; content tokening scheme; potential truth/satisfaction values; and type of extensional content.

Table 6: Winning & Bechtel's (2016) five dimensions of interpretation of information

Facet of information	Examples
Type of intensional content	Conceptual / non-conceptual, linguistic, experiential, propositional, imagistic, symbolic/non-symbolic, analog / digital
Type of vehicle	Representation (understood various ways), affordance, mere carrier, brute storage
Content tokening scheme	Causal covariation; statistical; biosemantics; teleosemantics; conceptual role semantics; Kantian categorical imposition; convention/stipulation; gestalt structuring; natural vs nonnatural
Potential truth/satisfaction values	None; true only; true or false; satisfaction (understood various ways); other possibilities
Type of extensional content	Actual objects; possible objects; rigidly designated objects; real patterns; facts; relations; events; property instances; tropes; states of affairs; abstracta

What is more important here than the contents of the table, is to recognize that the variety information offers, is precisely its key strength. That it can in principle be directed towards actual or possible objects, be about relations, events, or states of affairs, or have different kinds of truth or satisfaction values, and so on, is what makes it the only possible vehicle for experience. And even noting that information exists in a relationship, it yet seems to retain a curious autonomous status. The crucial thing to understand about information, is that it seems to arise in interaction (Roederer, 2003) in a way such that there is a peculiar gap between the information and the matter-energy that acts as its necessary substrate.

Anthony Wilden (1987, 71) defines information by contrasting it to simple matter-energy. For him, "Energy is the capacity of a system to do physical work. Information is the capacity of a system to do logical or structural work – its capacity to organize matter, energy, and/or information in ways not found in ordinary physical or chemical systems". Further, he writes, "Information in the simplest sense is a pattern of variety carried by a matter-energy marker or medium … [this] Variety has no intrinsic sense, meaning, or signification. For a pattern of variety or diversity to be acted on as information, it must form part of a coding system in a context. It must be part of a sender-receiver relationship organized by a goal or goals". In the context of this thesis, this definition is one that can really do work for us in advancing our understanding of the relationship between matter-energy, information, goals and by extension, agency. First, notice how this decouples the term information from its' material substrate. Information can't be thought of purely in terms of order and entropy, as in Shannon's (1948) model, simply because even pure noise, noisily distributed, can turn out to be information about something, as Deacon (2013, 382-382) notes through the example of the discovery of the cosmic background noise thought now to be evidence

for the Big Bang. Thus, information isn't necessarily in the signal, but in its' meaning within a context – for example theories and goals – of an observer. This is the critical point: the information being *about* something need not be based on any intrinsic properties of the signal medium (Deacon, 2013).

In colloquial use, the term information is used as a mass-noun to indicate for example the stuff that runs through our computers and smartphones, but also that which *informs* us, in the sense of "John gave me a vital piece of information for my thesis" (Adriaans, 2013). Gregory Bateson (2002) defined information as the "difference that makes a difference". Another way to put it might be that information has the character of "aboutness" to it. The vital piece of information for my thesis is about something which makes particular sense within the context of the thesis. A gene in a DNA molecule is "about" coding for a protein, or regulating other genes, which makes sense within the context of the organism. In this sense information is intimately tied in with teleonomic and teleological systems. On the other hand, our information age is grounded in Claude Shannon's 1948 mathematical theory of communication, but it should be noted that the theory was aimed at the engineering problem of noise: "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is *they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem."* (Shannon 1948, 1, emphasis added). Notice here that the semantic aspects are irrelevant given that the communication occurs between human beings, but as we move towards autonomous artefacts, the semantic aspects can no longer be treated as irrelevant to the engineering problem. They must somehow be turned into engineering problems that can be solved, not just assumed or removed from the equation.

5.5 Summary

The purpose of this section was to broaden the discussion to include the philosophical discourse around multiple realizability, supervenience, emergence, and forms of determination. Here we took a different approach to intelligence from computations. We tried both to affirm the autonomous existence and peculiar characteristics of the mental. But we also took a naturalistic perspective on intelligence and autonomy as particular forms of determination, predicated on information and exhibited in action. The key term was, as both cognitive science and AI would surely admit, information. But information as we conceptualized it was far wider and more complex category than what mere computations over representations might suggest. It is indeed the key, but information, specifically the semantic aspects of it, which we are ready to call informational proper, remains somewhat elusive. What is particularly problematic is whether semantic explanations (as causes of behavior) can be subsumed under a causal-formal explanation.

Nonetheless, the foregoing discussion has sharpened for us the need to clarify and connect the notion of information, emergent but distinct from concrete physical systems, as a necessary requirement for the instantiation of proper autonomy and intelligence in an artefact.

For the final section of this thesis, we turn to design in order to return the discussion back to technical artefacts and contextualize the discussion within the general theme of this thesis.

6 DESIGN

This thesis is ultimately about the design of autonomous technical artefacts *in general*. This means that the questions we ask and requirements and functions we define will be aimed at informing and supporting any similar pursuit. The purpose of this section is to first outline some characteristics of design, and second to try and tie some of the ideas presented in the previous sections into design thinking.

The general problem of this thesis is to *characterize the problem*, defined as the "distance or *mismatch* of the prevailing state and the state reflected by the goal" (Leppänen, 2005), in the attempt to design autonomous technical artefacts. The previous section was about identifying a mismatch between the concepts and tools afforded by AI and the goal of genuine autonomy. This section will be about taking constructive steps towards the general direction from which more suitable concepts might be developed.

6.1 Autonomous Ships as a Design Problem

Engineering design problems typically have two notable characteristics: they are open-ended and ill-structured. This means there are many possible solutions and the process by which solutions can be found is not always amenable to structured routine methods (Dym & Brown, 2012). Given some requirement, there is a space of reasonable technical solutions, but this space is difficult to circumscribe in any exact way. The space of possible objects can be quite large, some of which might not even satisfice as a solution and thus strategies of reducing this space are required (Chandrasekaran, 1990). Given the task of hammering a nail, things like hammers and shaped rocks occupy this space, televisions and chicken eggs do not. Given the task of attaching two pieces of wood, nails, hammers, but also glues and ropes enter the scene. This means that a significant part of the design process is the specification of requirements, functions, criteria, constraints and transformation of those into design problems. This is important, because without the right understanding of what is to be achieved, the fundamental open-endedness and ill-structured nature of design problems can emerge and cause havoc for the design process. This is a question also of quality: the more rigorously, widely, and specifically the requirements are represented, the less likely it is that corners will be cut or old solutions applied where new solutions are needed. Furthermore, proper identification and conceptualization of requirements shows the extent to which the design problem is either structured (solvable using standard techniques), semi-structured (standard solutions are available only to an extent); or unstructured (or wicked problem, which does not fit a standard model) (Leppänen, 2005).

As we have seen, perhaps the most significant aspect of the design problem of autonomous ships can be captured by saying that task for engineers is to replace human cognitive, perceptual, and motoric processes (thought of as semi-integrated in action) from the immediate vessel. This task can be approached from either the direction of devising technical solutions to particular scenarios, hoping to that they scale up with moderate amounts of tweaking, or by attempting to actually mimic or create artificial intelligence of a more general kind. Of course these are not necessarily opposed but complementary approaches. Our view is that proper autonomy entails the instantiation of general capacities in the artefact. This further means that the problem is essentially a wicked one, or at best semi-structured, meaning that we are still at a level of understanding whereby we do not necessarily know even what the right questions should be (Fodor, 1985).

Having stated the problem in a general way we can ask the question of the kinds of systems that can satisfy those requirements. Following our analogy before, what kinds of systems can conceivably occupy the space circumscribed by the requirements for proper autonomy? An approach closely aligned with this is more explicitly about concepts and theory languages. Namely, are we understanding the problem correctly such that a solution is, while perhaps difficult, still at least in the direction the questions point to? Do we have the conceptual tools by which to begin approach or even formulate the problems? Or taking another perspective, are the concepts currently available misguiding the search for solutions, or misrepresenting the problem?

6.2 General Introduction

Design, in the general sense of tool-creation and art, is a human universal. Along with language, it is what distinguishes man from other animals. The capacity for abstract thought, which they all presuppose, or perhaps enable, or are facets of, is what has enabled man to colonize essentially the entire terrestrial planet, with relatively small hard-wired biological adaptation to different environments. The ability to craft artefacts like clothing, weapons, and shelter enabled even early man to survive in quite different types of environments.

But what is design, exactly? In addition to being an adjective, design is often used both as a noun and a verb, to denote both the design artefact and the process by which it is created. To clarify this distinction, we will use the terms 'design processes' and 'artefacts' to separate the two. As our general topic is autonomous technical artefacts, our focus here will be on engineering design and here, *an artefact* or a *description of it* is the typical outcome of a *design process* (Kroese, 2013). What we call '*a design*' is, to use Kroese's (2013) "thick sense": a "description of a teleological arrangement of physical parts that together realize a function". That is to say, to achieve something, to realize some end, is why we create artefacts. Put another way, if an ontology is a way of carving up the world at its joints and an attempt to identify what Mario Bunge called "the furniture of the world"

(Bunge 1977), then what design does in a sense, it *changes* the ontology of the world: the proper *furniture* of it (Kerr, 2014).

Artefacts can be characterized by their functions and purposes, as well as the organization of their inner environment (Simon, 1981). A classic thermostat can be characterized by the function of maintaining a steady temperature within a room, and its' inner organization consists of a bimetallic strip that expand and contracts in response to temperature, which either closes or opens a circuit, which turns the heat on. As the temperature rises, the strip contracts which turns the heat off, which barring any other circumstances, most likely returns to a temperature which causes the strip to expand again, and the process repeats. This recursive feedback loop between the artefact and the environment in its relationship to human needs defines the artefact. The thermostat could also be digital, connected to a temperature gauge, without losing its' essential functionality – perhaps even enhancing it. Thus, a given design goal (such as maintaining the temperature in a room), can be reached by many different types of design. Typically, this kind of design thinking falls within the domain of engineering design. In the case of the thermostat, the variable of the environment to which the artefact is designed in relation to is the temperature.

But it would be a mistake to assume that in human use, a function is something static. In some sense, the function of a ladder might be to enable a person to climb up vertically, and this function may be what the designer has in mind. But when used, a ladder might equally well be used to cross a horizontal space, say over a stream of water, or two buildings. On the other hand, you couldn't very well steer a ship with a ladder, or at least expect the ladder to react dynamically to situations. Indeed, it is this very capacity of all human beings to see objects in terms of *some* goal that makes design, as a human capacity, possible at all - while at the same time making functional descriptions somewhat slippery. Nonetheless, function can't be dispensed with, since it is what makes objects *objects* at all, let alone tools, rather just physical things (Peterson, 2013). In fact, even that is a stretch, since something like a "purely physical" description of a hammer already presupposes a representation for some end, to strip it of its' functional connotations for example. As far as human thought is concerned, there are no 'pure' descriptions, only silence. The object's participation in a human mental representation, characterized by both negative and positive selectivity (Saariluoma, 1992), is what grants it object-status. This is more pronounced in the case of artefacts, fashioned and used as they are within the domain of human affairs, social and cultural.

A further difficulty in nailing down a formal description of design is that form (or structure) does not strictly logically speaking follow function (Kroese 2013). This is because of the shifting possibilities of functions in human use, and because many different forms can achieve a single function. What this shows is the somewhat open-ended and creative nature of design and especially novel design or innovation, but also the possibility for re-appropriation of existing technology. Indeed, an innovation is often a confluence of other innovations (Saariluoma, Hautamäki, Väyrynen, Pärttö, & Kannisto, 2011). That many forms

can achieve a single function is clear, but it does not mean anything goes. In the ladder example, what both functions presuppose is a certain rigidity, the ability to withstand the weight of a person. That this is the case is what the user will at least tacitly assume or hope, although of course such assumptions may turn out fatal. On the other hand, we would hardly assume a ladder to be able to perform the function of a pocket calculator. The form of a ladder has in typical human use a tacit range of possibilities, which we can in some sense intuit.

A design process is typically a long series of questions covering (tacitly or explicitly) all the aspects that make up the final product (Saariluoma 2009). The first questions a designer must ask are about what is required to achieve the functionalities that are being sought. This, as mentioned previously, is in our case the transfer of all relevant tasks, functions, and capacities from the domain of man to the domain of the vessel. This in turn is transformed into a design problem.

The formal definition of a problem is to reduce the difference between the current state and goal-state of the system (Newell & Simon, 1976; Leppänen, 2005). One way designers attempt to reduce this difference is by generating alternatives and evaluating/testing them against requirements and constraints (Hevner, March, Park, & Ram, 2004; Lawson ,1990). The generation of alternatives does not, however, come from nowhere. It is presupposed by the requirements and functionalities being sought for on the one hand, and on the other, the tacit and explicit knowledge and skills of the designers. In other words, both problems and solutions are constrained and made visible by the concepts employed by the practitioner (Saariluoma, 2009). A programmer, an engineer, a psychologist, or a lawyer (ap)perceive a thing under consideration quite differently based on their conceptual structures and learning (Floridi, 2017).

In addition to their characteristic artificiality, all design processes are the result of human thinking. According to Lawson (1990) all design involves "a highly organized *mental* process capable of manipulating many kinds of information, blending them all into a coherent set of ideas and finally generating some realization of those ideas." (see also Dym & Brown 2012). It is strictly speaking wrong to assume that design can, or should, be divided into sharp categories without interaction and overlap. Most obviously, a technical artefact consists of many facets which require different sets of understanding: a car has a motor and it has an aesthetic, but an aesthetic should usually accord with the laws of aerodynamics: many forms of understanding converge in engineering design (Vincenti, 1990). On the other hand, many innovative ideas have not come from people deeply vested in the domain of the invention. For example, Lawson (1990) lists some: the ballpoint pen was invented by a sculptor, the parking meter by a journalist, and the automatic telephone by an undertaker. From this he draws the obvious conclusion that we shouldn't classify design by its end-product lest we wish to unnecessarily straightjacket designers, and direct their mental processes towards predefined goals, if what is needed is innovative design. Yet, design is not, nor should it be, completely open-ended. The trick is how to define requirements in such a way that ingenious designers can connect the dots with whatever means and concepts they have available.

This is the meta-justification for the attempt of this thesis – to provide a contribution to the design of autonomous technical artefacts from the perspective cognitive science (broadly understood).

6.2.1 Science and Technology

In the early 1960s, authors began to articulate the nature of technology and technological progress, specifically with respect to its' relationship with science (Franssen 2013). Henryk Skolimowski (1966) and Herbert Simon (1981) noted how the essential difference between science and technology is roughly one of values: science wants to understand existing phenomena and advance understanding of them; engineering design seeks to fashion artefacts that accomplish some defined purpose. The two domains of human endeavor have obvious connections. Sometimes scientific discoveries find application in technology, sometimes the goals of technology advance scientific understanding. Indeed, as Bunge (1966) notes in the same issue as Skolimowski (1966) (while defending the notion of technology *as* applied science), technology is differentiated from *craft* precisely by the application of scientific knowledge in solving technical problems and finding solutions. Yet, as Franssen (2013) notes, there is a danger of missing the most central distinguishing element of technology if one focuses on its' scientific aspects, namely design.

Design brings in the notion of possibilities. We might say that the physical laws underpinning nuclear bombs would have existed irrespective of the human goal of creating weapons of mass destruction, but the objects themselves would not. Nor, for that matter would Plutonium, as it is not an element found in nature (Skolimowski, 1966). Thus, technology and design are areas of life where forms of logic such as modal, praxeological, and deontic apply (Saariluoma, Canas, & Leikas, 2016). What is important to note, is that while design as an activity is through and through teleological in nature, the concepts *by which* it operates, or attains its goals may not, indeed often are not teleological as such. Explanations of this sort, as we have discussed before, are perhaps best captured, as von Wright noted (2004), via practical syllogisms. Namely, the major premise of the syllogism is some envisaged end state, the minor premise relates some action as a means to that end and the conclusion is simply to use the means to reach the end. It is hardly any surprise that in normal human affairs the stability afforded by causalistic and simple-mechanical approaches is very useful. Perhaps it is no surprise either, that they are unable to approximate the capacity in service of which they are used: the higher cognitive processes of man.

6.3 Design Phases

Quoting the American Institute of Architects, Floridi (2017) outlined a rather uncontroversial, phased account of design:

Phase 1: Originate This is the thinking phase in which one realises that something new (in our terminology, a new system) needs to be built in order to satisfy a particular purpose. It is the "needing" moment of the project.

Phase 2: Focus This is the phase in which one defines the system's requirements (these are broadly understood as scope, features, purpose, or functionality, more on this presently) that the system must have in order to satisfy the purpose. It is the "vision" moment for the project.

Phase 3: Design This is the phase in which one models the system's requirements. It is the "shaping" moment of the project.

Phase 4: Build This is the phase in which one constructs the system. It is the "making" moment of the project.

Phase 5: Use This is the last phase in which the system is finally available and starts satisfying its purpose. It is the "testing" moment of the project.

Of course, the reality is far less linear, as Floridi (2017) himself notes, and this account but one among many (Lawson 1990), but it satisfies its' purpose for now. It seems like at present, autonomous ships are somewhere between phases 2 and 4. Our interest in this thesis can be anchored between phases 2 and 3, namely, the identification and discussion around requirements entailed by autonomy, as we have defined it. Furthermore, to consider the conditions of feasibility and possibility implied by the requirements vis-à-vis machine intelligence. This line of thinking was outlined by Floridi (2017) in his notion of the logic of design as a logic of requirements.

6.4 Logic of Requirements

Luciano Floridi (2017) has outlined a logic for design built around *requirements* and *conditions of feasibility*. The requirements come out the view of design as more or less open-ended, with many different systems being capable of serving a single requirement (Dym & Brown 2012, Simon 1981). Thus, through the language of requirements, the approach does not unecessarily or unrealistically constrict the design processes. This he contrasts with *conditions of possibility* which he traces to Kant's transcendental logic. This approach looks at systems and tries to understand what brought them about. As such, it is close to the methods of natural science. Thus, Floridi's logic for design is oriented towards the future, towards feasible possibilities, whereas the scientific stance is oriented towards what is, and more generally or ideally, that which is all-encompassing and eternal, such as natural laws. The difference between these two approaches is what Saariluoma (2010) identified with the scientific and design stances (see also Skolimowski 1966 and Simon 1981). These two stances are by no means contradictory, but complementary. The search for a solution to a design requirement can be associated with the relevant scientific understanding. This is likely to be the case in any design

goal that is challenging enough and which likely does possess conditions of possibility that are at least not straight-forward and obvious. Autonomous ships as a case certainly qualifies. However, before a design problem can be "enriched" by scientific understanding, it needs to be identified, and these emerge ultimately from the requirements placed on the artefact, which in turn must be properly conceptualized.

Let us now recall the vision set out for autonomous ships in the AAWA project. Recall that we noted that the goals are relatively realistic and modest compared to what full autonomy actually entails. Indeed, the project's idea of ship autonomy as a *variable* depending on context and tasks rather than as a *constant* leaves the development somewhat open and places the most important requirement on how this variable is handled in real-life situations. In other words, how the shore control center can reliably and knows when to take over. Thus, the functionalities of the shore-ship link and center itself forms the superordinate requirement for the system at this stage whose importance will vary as a function of the level of autonomy achieved in the artefact (as a constant). Of course, this is still far from autonomous, rather more like unmanned & remote controlled combined with some capacities for sensing the environment, avoiding simple collisions, and maintaining a preset course.

Given that our interests are in the long-term (20-50 years) prospects of autonomy, we must however look beyond what is currently envisaged.

6.4.1 Dimensions of Autonomy as Requirements

In the first section, we presented the dimensions of autonomy as identified by Williams (2015). Let us consider now them as requirements. Williams (2015, 54) summarized the key dimensions of autonomy for technical systems as follows.

Table 7: Key dimensions of autonomy (Williams, 2015)

Autonomy dimension	Definition
Goals	An autonomous agent has goals that drive its behaviour.
Sensing	An autonomous agent senses both its internal state and the external world by taking in information (e.g., electromagnetic waves, sound waves).
Interpreting	An autonomous agent interprets information by translating raw inputs into a form usable for decision making.
Rationalising	An autonomous agent rationalises information against its current internal state, external environment, and goals using a defined logic (e.g. optimisation, random search, heuristic search), and generates courses of action to meet goals.
Decision making	An autonomous agent selects courses of action to meet its goals.
Evaluating	An autonomous agent evaluates the consequences of its actions in reference to goals and external constraints.

(continued)

Table 8: Key dimensions of autonomy (Williams, 2015) (continued)

Adapting	**An autonomous agent adapts its internal state and functions of sensing, interpreting, rationalising, decision making, and evaluating to improve its goal attainment.**

What is key to note, is that the dimensions form a whole. They may be thought of as "layers" or "modules", and indeed are likely be so construed given the way machines are constructed and operate, but as far as autonomy is concerned there is, for example, no decision-making without rationalizing or interpreting, or without goals.

What is needed, is an ontology for the design of technical systems that can achieve a mapping between the domain of man and the domain of technical artefacts, and this quest may involve a reconceptualization of how this mapping is to be achieved, if we are to follow the critiques outlined in the previous sections.

6.5 The Artificial

What all design shares in common, from furniture to intercontinental ballistic missiles, is their characteristic artificiality (Simon, 1981). The basic difference between a natural and an artificial system is that the latter is the result of human intentions. As articulated by Herbert Simon in his "Sciences of the Artificial" (Simon, 1981), the peculiar characteristic of the artificial is that it embodies both human intentions (goals, needs, functions) and the laws that govern the natural world which are, of course, nowhere absent. For Simon (1981), the natural is marked by a sense of necessity, whereas the artificial has an air of contingency[5]. The operating principles of cars for example are not removed from general laws of nature but the laws are, so to speak, contained and directed within a structure developed by human ingenuity to serve human purposes, i.e. locomotion. It is the harnessing of the regularities in nature within a form that satisfies some purpose. A machine for example is often a system by which energy can be transformed to do work. A chair has very different kinds of purposes, but nevertheless it remains embedded in human needs and thinking.

This definition of the artificial ties in interestingly with artificial intelligence. Given Simon's (1981) acknowledgement that artificial systems are the result of human intentionality, it follows that such systems have by default no intrinsic intentionality. There is no clear sense in which the sense of striving which we associate with life could issue from a machine if we admit to a fundamental difference between the two (Rosen, 1999; Thompson, 2010). In other words, even seemingly intelligent software is basically just the displaced intentions of the prescient programmer (Deacon 2013, 100). The operative word is "by default".

[5] It might be added, that life itself is marked to some degree with an air of contingency, given the sense of finitude and precarious striving between life and death of the individual life forms that constitute the whole.

Human beings are "by default" intentional agents. Machines are "by default" the displaced intentionality of human beings, and have no intrinsic intentionality. But, of course, the goal of autonomy and perhaps indeed a requirement for general AI is to create systems that *do* have intrinsic intentionality. Although biologists (Mayr, 1974, 2014) have long since expunged their field of teleology and essences, it may be that some of the most interesting properties (such as teleology and intentionality) and the way in which they may connect with not essences, but shared properties in common to all life, are thus left unattended (Jonas, 2001; Rosen, 1999). It may be of interest to note, that the primary tool by which teleology was ousted from nature was the theory of evolution (Mayr, 2014), but that it provides exactly *zero* insight into the emergence of life from inorganic matter, only its subsequent variations (Deacon, 2013, p. 138), and the cold fact remains that while theoretical advances have been made, so far no one has succeeded in creating life in a laboratory (Mayr, 2014).

6.6 Concepts and Languages

Two forms of human life, and the differences and connections between them, are relevant to understand in the context of this thesis: science and engineering design. More specifically, the science of mind and intelligence, but it requires a separate treatment after the initial clarification is done. The difference, as noted by Checkland (1994) is roughly one of values: science wants to understand existing phenomena and advance understanding of them; engineering design seeks to fashion artefacts that accomplish some defined purpose. Technology, understood as a merger of science and engineering, is a very powerful technique indeed, and is perhaps a central prerequisite for modern civilization. The fact that they *can* have such an interplay is predicated on the fundamental similarity of the languages by which they explain and accomplish their respective tasks: causal explanations. Scientific explanation, as defined by Hempel (1942) in the deductive-nomological model (for its' applicability to human affairs see Saariluoma, 1997; von Wright, 2004), has two major constituents: the explanandum and the explanans. The essential task is to map on the target phenomenon to be explained (explanandum) some laws, sentences, or equations that account for the phenomenon (the explanans). It is deductive in that the explanandum should follow as conclusion from the explanans. It is nomological because the explanans should have, or contain, lawlike regularities (greek expression -nomos means lawlike). An explanation of this form combined with engineering design can be called technology: the artefact is not a product simply of craft, but we understand why it works, and also why it does not based on general laws of nature. It should be noted, of course, that this connection should not prompt us to discard the important difference illustrated before, and the fact that both purposes for engineering design and purposes for science do not land in the explanatory methods they employ, but in human needs, culture, social practice, and individual minds (Saariluoma, Canas & Leikas, 2016). Some science and some engineering can enjoy a rich dialogue.

But it is not clear whether the language in the sciences of mind, and engineering design can enjoy similar interplay if the critiques of AI outlined before hold any truth. It should be noted, that it is precisely this interplay, or the attempt thereof where engineering design meets cognitive science and artificial intelligence. But what has vexed cognitive science, artificial intelligence, and communication theory has always been meaning and significance. It is not clear how the translation from meaning and significance to matter in motion (Dennett, 1986) should proceed without something essential getting lost in the way such that it needs to be injected to the theoretical or physical edifice from the outside, ad hoc so to speak. Again, this marks the difference between autonomous and automatic. Now it is not necessarily the purpose here to argue in some general sense against the possibility or efficacy of causal explanations as explanations for some facets of the mental. The point, as a distinction borrowed from Burge (2007) points towards, is that there is a difference between causal *antecedents* and causal *consequents*. In the sense relevant here, this is not to be thought of as some kind of input-mechanism-output system in the flat sense which easily collapses into mere automation, but by thinking about the causal antecedents, the necessary conditions for mind to emerge from matter in motion, but which thereafter is peculiar characteristics of its' own – even as it remains firmly tethered to its' causal ground, call it the body or the brain. The way to think about this is via the special qualities of minds and informational structures in general: their connections are best described as logical, or syntactic, and sometimes meaningful or semantic. The problem that vexes the causal picture of psychology, is that there seem to be no "gaps" into which some non-physical or non-physiological aspects might inject themselves (Burge, 2007), the causal picture is complete and gapless. What is the weight and dimensions of thought such that it can bear weight and cause events in the brain? What of information? Such questions sound non-sensical, like asking of the color of number three (Wittgenstein, 1994).

Insofar as we acknowledge actionistic language as a valid theoretical construct in its' own domain (most human affairs, the philosophy of action, much of psychological explanation), and insofar as we grant it as pointing towards something essential for the attainment of autonomy, we come upon a deep chasm between two theory languages: the causal language of engineering (Pahl et al., 2007) and the language of action, with its' teleological and intentional characteristics. To be implementable, a function needs to be rigorously defined for the engineer. We can simply *ask* a person to fix breakfast tomorrow morning because we ourselves will be busy replacing the tire on our bike, but such a description, while perfectly functional in human affairs, simply can't do work in engineering terms. An engineer might translate such language to some materially-causally feasible form, but in and of itself it will not be sufficient. The question becomes then how to achieve a mapping between two distinct knowledge patterns? The general approach of cognitive science has been to map mental processes to general descriptions of information processing which acts as the abstraction and bridge between man and computer.

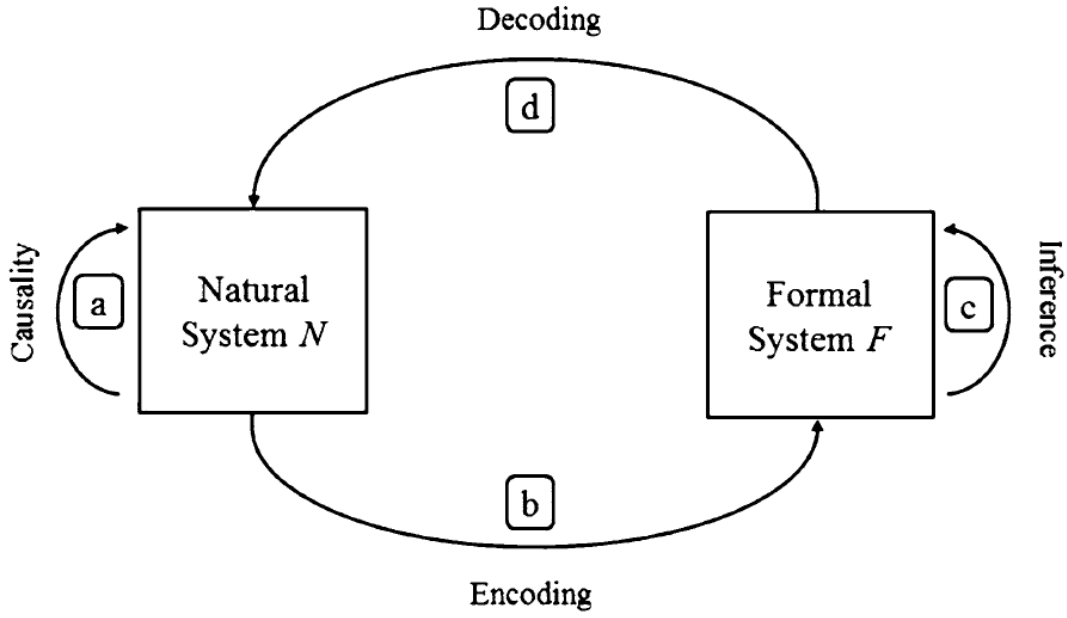
The tension between the intentional actionistic language of everyday human affairs, and the language of engineering design is telling of a more general tension to do with any translations of phenomena into a theoretical framework: the case of phenomenological phenomena simply makes the point more vivid. Insofar as the advance of scientific understanding is a conceptual affair (Kuhn, 1970; Saariluoma, 1997), it is an attempt to map, or translate, some source phenomenon into some logical framework. Insofar as no map is the territory it represents, *nor should it*, there are always patterns that get left out, either by design or by other reasons. The triumphant mapping of maps to yet more maps, landing perhaps in physics, would the goal of a reductionistic analysis. This may be with regards to any topic, but here we are of course thinking about the mapping of mental processes to physical events in the body and in the environment. The autonomy of a science (of the mental, say) turns on the possibility of such a reduction. Or one might even turn the tables and ask whether the physical can in some sense reduce to the mental, although this is perhaps a less typical direction of reduction. Two central problems posed by the apparent qualities of the mental pose a significant challenge to a reductionistic analysis. One is the holistic nature of the mental. The other is the wholesale disappearance of subjective qualities, arguably the mental, that ensues from reduction to a causal theory language. It seems in no trivial sense that something essential is lost during such a translation. It may follow, that if something essential is lost along the way, the return trip from matter in motion to meaning and significance (Dennett, 1986) will not reinstantiate the central features of the mental, but mere simulacra. This, if true, is a severe limitation on artificial intelligence, and on any cognitive theories that resemble the approach for that matter.

We are ready to admit both the validity of mental, viz. actionistic, intentional, and teleological, explanations in human affairs, and also ready to admit its' insufficiency with regards to engineering science and therefore artificial intelligence (understood as the attempt to recreate the phenomenon of the mental). We will also note, that cognitive science was and is the attempt to postulate explanations of the former kind (often called folk-psychological) into an intermediary, cognitive, theory language that would be more amenable to a scientific treatment (Frankish & Ramsay, 2012). But just as the language of ordinary human affairs will not suffice or by itself do work in the context of engineering design, a different, perhaps more pernicious kind of error is one which *seems to*, *kind of*, do such work – especially if it is at least biased, or prepared to, eject meaning and semantics from the explanatory framework as *meaningless.* The wrong type of explanation can't hope to recreate the causal entailments which we will assume exist as antecedents of the mental. While descriptive models can be harnessed to do some work in the context of artificial intelligence, it is reasonable to ask whether such a model has any real hope of advancing beyond empty mimicry or expanding our understanding of the mental in general. It is precisely in the context of artificial intelligence or genuine autonomy that explanations akin to Ptolemaic epicycles of planetary motions that are only descriptive but do not penetrate deeper into the reasons why (Bunge, 1979) will not do the work required.

Our position is therefore quite a conservative one. We simply acknowledge the existence of the mental with its' peculiar abilities and characteristics and combine that with the traditional view that the existence of things have reasons, viz. causes, or forms of determination (Bunge, 1979). Our position here is that a mark of a science of the mental would either be able to show a principled way towards recreating it, or a principled reason against that possibility: it would turn the mental into a technological object rather than the current folk-psychological conceptions currently dominating the field of AI. Would it not be the case that in the event that humans succeed in creating actual minds in machines, there would at that very instant be born a need for an autonomous science of "machine psychology"? The difference would be that unless the thinking machines came about by some strange accident, then here we would know the causal antecedents of the mental in the machine, and yet, the actual subsequent psychological *space* would be in some sense unexplored territory.

6.7 Models and Modeling

The tacit assumption with any attempt to model a phenomenon is that it can capture in an *inferential* structure some parts of the *causal* structure (Rosen, 1999). Yet, the more autonomous and/or complex the target phenomenon the less possible it becomes to capture the actual entailments within a (typically) computer model. This is not necessarily a reflection of our ignorance of the causal connections as such, although it is also that, but of the limits of formal models, and by extension traditional machines including computers and therefore AI. As noted by Rosen (1999) and reiterated by Schierwagen (2012), the essential relation between a model (or a blueprint) is to capture the causal (broadly understood) relations in a natural system within a formal system in the form of inferences. See image below (from Schierwagen, 2012):



The relationship between models and natural system (Schierwagen, 2012)

The relationship between the two systems are of encoding (b) and decoding (d). The fundamental problem is that only the simplest natural systems and the most general patterns afford such analysis. Autonomy, and indeed the mental and its' living substrate are far more complex. The problem with this is two-fold. Computers function by way of formal systems. This foundational limitation, which we have identified before, means that they can, at best, approximate in simulation any process that has complexity as its' characteristic (Checkland, 1994). This problem relates both to the operating principles of computers vis-à-vis natural organic systems, but also to the situations and environments in which the AI system is to manage (Rosen, 1999). These two aspects, complexity in the environment and complexity in the system, may be related. On a general level, Simon's (1981) insight that the apparent complexity of man is but a reflection of the situations in which he finds himself rather than an intrinsic quality of the system itself is both right and may be wrong, or traffic within a notion of complexity that refers to complicatedness rather than complexity (Rosen, 1999). To be sure, the idea of a complex system, while important in many fields, seems to lack a concise definition applicable and accepted across different disciplines (Ladyman, Lambert, & Wiesner, 2013). Mathematical biologist Robert Rosen (1999) defined a system as simple if all its' models are simulable. Accordingly, a complex system is one which must have a nonsimulable model. Importantly, for Rosen, a simple system can be *complicated* without crossing the threshold to complexity. Furthermore, he held that the neat, tidy, orderly world of mechanisms is the world of the simple. This he further identifies with the formal system, nice and clean, but impossible to get out of via methods internal to the system. This is a fascinating insight, that has a direct family resemblance to the critiques of machine intelligence put forth by Dreyfus (1976, 2007) and Searle (1980, 1984, 1990).

Namely, the formal system is for Rosen, as it is for Dreyfus and Searle, a closed circle and an entity in it can't reason its' way out (consider also Wittgenstein, 1922). It is a purely syntactic system, without meaning and without relevance. And, it is the world of the machine, fully determined, only syntactic. Finally, whether systems that have hybrid characteristics, some neat and simulable and others not, and how those line up with properties we are interested in is a question difficult to prejudge (see Marr, 1990).

Simon (1981) saw clearly that human thinking is governed by heuristics and rules-of-thumb which sacrifice precision for usefulness and speed (see also Kahnemann, 2011). A typical heuristic for human thinking in the context of design is the so-called mereological and decompositional strategies that perhaps clearest in engineering design. Mereology is the study of parts, and decomposition is a related strategy of carving a system by its joints and analyzing them as separate components or modules (Bechtel & Richardson, 2010). Whether it is a viable strategy depends on what Simon (1981) called either near or full decomposability, which means essentially that the interactions within subsystems or modules are stronger than those between them. The problem is that this is the only way we can build systems and artefacts, we don't copy the bird in building an aeroplane because we can't copy a bird (Rosen, 1999). This is, of course, a problem only if the function we are attempting to capture resists mereological or compositional strategies and the property depends on such interactions as to make progress impossible without different conceptual tools. Recall for example our discussion on the modularity of mind (Fodor, 1985), which seems to indicate that even if some parts of the cognitive system may fall for a decompositional analysis, some of the most important and interesting may not, and furthermore, those may be absolutely crucial for true intelligence, and therefore autonomy. As the difficulties but also successes yielded by attempts at carving the cognitive at its joints seem to suggest (Janssen, Klein, & Slors, 2017) there is reason to believe that there is truth in both viewpoints.

Finally, we should keep in mind that a very basic problem lurks within the idea that a formal model could capture in an inferential structure the causal structures that underpin the mental. This is a slippery point, for isn't the mental, as we have suggested, precisely inferential and not causal, and shouldn't it therefore follow that the causal structure is not important? In the Artificial General Intelligence (AGI) community, for example, the received view is that creating AGI is "just an engineering problem" given that humans, that do display AGI, are just "particular configurations of atoms" and an exact copy, down to the atomic level, or the human brain in a digital simulation would be "an almost sure way to create AGI" (Goertzel & Pennachin 2007, pp 17). But this is too disembodied (Clark, 1997) a perspective for our taste. It seems likely that precisely by only projecting our imaginations on artefacts we fall short of achieving anything resembling intrinsic intelligence in the system, if we take seriously what our experience and common sense tells us, and take seriously the self-experience of life (Jonas, 2001) as an phenomenon curiously tethered to the body, emergent from it, and yet open to an informational aspect of reality which

one is tempted to acknowledge genuinely as 'a third realm' (Popper, 1979; Wiener, 1985). It is possible that no amount of simulation, without the right mechanisms that turn on the mental lights, will ever yield much more than complicated, but ultimately (not even) dumb software. The moral dimensions of even wanting to turn on mental lights in artefacts will have to wait for another occasion.

6.8 Summary

We have arrived at the limits of what this thesis can achieve. Our discussion has sought to identify that the requirements for full autonomy as predicated on genuine intelligence in the artefact require more than current understanding can achieve. This fundamental tension is thus at this stage of understanding a conceptual one, not yet a technical one. In order to have a chance at instantiating true intelligence, or true autonomy in a technical artefact, we must have the proper conceptual tools for that endeavor. What such concepts might entail one can only guess. What is obvious, is that mental contents and meaning are real fixtures of reality. So far, we have only the slightest glimpse of how we might properly explain how they come about from mere matter in motion. If such an explanation is to be found, it needs to take seriously the various properties of reality, and seek to identify how higher-order constraints, information, come about in physical systems, and how to properly conceptualize it. What is needed is an ontology for the design of autonomous technical artefacts that could, without losing what is essential on the way, achieve a mapping from the broadly cognitive domain into the technical domain, and may entail a reconceptualization of how the artificial is to be built, if built is even the operative word.

In the meantime of course, we need not despair. The technical systems have an autonomous mental component: human beings. Before a much deeper understanding of how autonomy comes about in us has been achieved, the human will remain a necessary component of technical systems.

7 CONCLUSIONS AND DISCUSSION

We began this thesis by approaching the question of autonomy in the context of maritime vessels. We noted that if autonomy is the requirement, then intelligence of one form or another is the solution. Intelligence understood as a catch-all term here, and including perceptual capacities. It therefore follows that the level of autonomy is in some sense a function of the intelligence embedded in the technical system. Our brief overview of the various problems involved vis-à-vis the technical systems on offer concluded that, first, the level of ambition in the Design for value project is at a realistic level, and that given that anything like full autonomy is still somewhere in the future, the primary term should be unmanned, and the primary solution should be a shore control center. Ship autonomy is a variable rather than a constant, and it will likely remain that way for some time. Significant focus should be placed on how, for now, the capacities formerly exercised on the bridge and the deck can and should be transposed into the shore control center, and presents a necessary research field for the upcoming years. The development of autonomy in ships can proceed in parallel, but should not be deployed or relied upon before the remote control issues are settled. Of course, if the ship if unmanned, it needs to have robust solutions for handling situations if the link is severed or uncertainty exceeds some threshold. Much work remains to be done.

The technical difficulties involved open up the path towards a discussion on conceptual issues buried within the question of autonomy, insofar as it relates to intelligence or what we might associate with higher cognitive processes in man. Namely, while there is no question as to the technical difficulty of achieving anything like general human intelligence in an artefact, it is also possible that those technical difficulties, in part, result from conceptual confusion. These questions are what preoccupied us over the latter parts of the thesis. The problems and questions that these questions opened up are many and interweaving and our discussion could but illustrate one possible path through this vexing territory.

We began with a discussion on the very nature of the digital computer, the most usual suspect as the platform on which artificial intelligence is to be instantiated. Following quite classical critiques of AI, we concluded that the question of meaning and semantic contents remains conceptually, in principle, outside the grasp of computers – understood as formal-syntactic systems – and that therefore, even their fundamental operational principles, namely computations, are to a non-trivial extent observer-relative. The problem can be conceptualized as issuing from a mapping of abstractions, which by definition leave contents behind, into a regular mechanism. There is no obvious way how such a procedure, especially given that it presupposes the jettisoning of contents, could by miracle come to intrinsically exhibit them.

This began to turn our discussion towards broader issues given that if computers do not suffice, but apparently we do, what is it about us that grounds this potential? This we attempted to approach via the notion of multiple realizability

by acknowledging the special qualities of the mental vis-à-vis the physical, and that the question is really about how we are to conceptualize the mental and the physical such that a justified plurality would be the starting point before any sort of reduction would even be envisaged. Our approach was a kind of scientific realism, namely, the very existence of the autonomous science of cognitive science and psychology, not to mention our ordinary experience and communication, affirms an autonomous status to mental events and causes. They are real and efficacious. This opened up the possibility of examining the questions from two different points of view: the operating principles of the mental as a form of determination on the one hand, and on the other, its' causal antecedents. That is, by acknowledging a plurality we can avoid any foregone reductions and yet examine the antecedent, roughly speaking physical, causes of the mental which are admittedly its' sine qua non. This led us towards an assessment of information as indeed the crucial notion, but that our understanding of it is on par with our understanding of how semantic contents come about in natural systems, let alone artificial ones.

What our discussion sought illustrate, is that a plausible conceptualization of mentality should take seriously the forms of determination by which it emerges from physical substrates, and its' subsequent qualities as a form of determination itself. One approach is to think of information and the mental as ways in which lower level forms of determination are captured and constrained within patterns and structure. The major question remains, and is not obviously explained by the ways sketched in this discussion, of how mental contents come about. It is possible however, that to seek answers to these questions one should go down to the very nature and origins of life itself with the assumption that something like meaning, dim as it may be in the beginning, is a necessary corollary to life.

For the final section we sought to land these abstract considerations in the design of autonomous technology, the general topic of this thesis. Understanding the nature of design as an exploration of possibilities, we sought to anchor the discussion in a framework of requirements, namely, that the dimensions of autonomy illustrated in the first section are indeed the requirements for it. The problem, as the middle part of this thesis sought to show, is that they are somewhat descriptive of what it entails, and do not as such translate into a language useful for engineering design. Thus the discussion around AI and multiple realizability can be seen as a more broad illustration of a requirement for new conceptual tools and language that genuinely tackle the problems that proper autonomy entails. Here engineering design, AI, and cognitive science enter common ground, and a genuine solution from any direction should benefit each. It is possible however, that an engineering design that actually achieves proper autonomy would require a major conceptual shift away from the typical mereological summing of parts and modules towards a deep integration and forms of holistic determination. But as the history of AI and computer science show, placing these theoretical considerations in the context of actually attempting to achieve them

has immense potential for the advancement of human understanding. The importance of negative results should not be underestimated.

REFERENCES

Abrahamsen, A., & Bechtel, W. (2012). History and core themes. In K. Frankish, & W. M. Ramsey, *The Cambridge Handbook of Cognitive Science* (pp. 9-28). Cambridge, New York: Cambridge University Press.

Adams, S., Arel, I., Bach, J., Coop, R., Furlan, R., Goertzel, B., . . . Sowa, J. F. (2012). Mapping the Landscape of Human-Level Artificial Intelligence. *AI Magazine*, 25-42.

Adriaans, P. (2013, July 12). *Information*. Retrieved from the Stanford Encyclopedia of Philosophy (Fall 2013 Edition): https://plato.stanford.edu/archives/fall2013/entries/information/

Aho, A., & Ullmann, J. (1994). *Foundations of Computer Science.* W. H. Freeman. Retrieved from http://infolab.stanford.edu/~ullman/focs.html#pdfs

Ahvenjärvi, S. (2016). The Human Element and Autonomous Ships. *TransNav The International Journal of Maritime Navigation and Safety of Sea Transportation*, 517-520.

Anton, C. (2011). *Communication Uncovered: General Semantics and Media Ecology.* Fort Worth, Texas: Institute of General Semantics.

Anton, C. (2012, February 4). *Dialogue on Information, Science, & Levels of Reality (Susskind's Hologram too)*. Retrieved from YouTube: https://www.youtube.com/watch?v=0o2FiSz-57k

Bateson, G. (2002). *Mind and Nature: A Necessary Unity.* Cresskill, New Jersey: Hampton Press Inc.

Bechtel, W. (1994). Levels of Description and Explanation in Cognitive Science. *Minds & Machines 4*, 1-25.

Bechtel, W., & Mundale, J. (1999). Multiple Realizability Revisited: Linking Cognitive and Neural States. *Philosophy of Science 66*, 175-207.

Bechtel, W., & Richardson, R. C. (2010). *Discovering Complexity : Decomposition and Localization As Strategies in Scientific Research.* Cambridge, Massachusetts: The MIT Press.

Bennett, S. (1984). Nicolas Minorsky and the Automatic Steering of Ships. *control systems magazine*, 10-15.

Bickle, J. (2016). *Multiple Realizability*. Retrieved from The Stanford Encyclopedia of Philosophy (Spring 2016 Edition): https://plato.stanford.edu/archives/spr2016/entries/multiple-realizability/

Bitbol, M. (2007). Ontology, matter, and emergence. *Phenomenology and the Cognitive Sciences*, 293-307.

Blanke, M., Henriques, M., & Bang, J. (2017). *A pre-analysis on autonomous ships.* Lyngby, Denmark: Technical University of Denmark.

Boden, M. (1990). *The Philosophy of Artificial Intelligence.* New York: Oxford University Press.

Boden, M. (2008). Autonomy: What is it? *BioSystems (91)*, 305-308.

Boden, M. A. (1987). *Artificial Intelligence and Natural Man.* London: The MIT Press.

Boden, M. A. (1989). *Computer Models of Mind.* Cambridge: Cambridge University Press.

Brentano, F. (2009). *Psychology from an Empirical Standpoint.* London and New York: Routledge.

Brooks, R. (2018, April 27). *[FoR&AI] The Origins of "Artificial Intelligence"*. Retrieved from Rodney Brooks - Robots, AI and other stuff: http://rodneybrooks.com/forai-the-origins-of-artificial-intelligence/

Bunge, M. (1966). Technology as Applied Science. *Technology and Culture Vol. 7, No. 3* , 329-347.

Bunge, M. (1977). *Ontology I: the Furniture of the World.* Dordrecht-Holland Boston-U.S.A: D. Reidel Publishing Company.

Bunge, M. (1979). *Causality and Modern Science.* New York: Dover Publications Inc.

Bunge, M. (2004). How does it work? The search for explanatory mechanisms. *Philosophy of the Social Sciences, Vol. 34 No. 2*, 182-210.

Burge, T. (2007). Philosophy of mind: 1950-2000. In T. Burge, *Foundations of mind* (pp. 440-465). Oxford New York: Oxford University Press.

Chandrasekaran, B. (1990). Design Problem Solving: A task analysis. *AI Magazine Volume 11 Number 4*, 59-71.

Chandrasekaran, B., Josephson, J. R., & Benjamins, R. V. (1999). What are ontologies, and why do we need them? *IEEE Intelligent Systems*, 20-26.

Chauvin, C., & Lardjane, S. (2008). Decision-making and strategies in an interaction situation: Collision avoidance at sea. *Transportation Research Part F*, 259-269.

Checkland, P. (1994). *Systems Thinking, Systems Practice.* Chichester New York Brisbane Toronto: John Wiley & Sons.

Clark, A. (1990). Connectionism, Competence and Explanation. In M. A. Boden, *Philosophy of Artificial Intelligence* (pp. 281-308). New York: Oxford University Press.

Clark, A. (1997). *Being there: Putting Brain, Body, and World Together again.* Cambridge, Massachusetts: The MIT Press.

Cross, N. (1999). Natural intelligence in design. *Design studies (20)1*, 25-39.

Cull, P. (2007). The mathematical biophysics of Nicolas Rashevsky. *BioSystems 88*, 178-184.

Davidson, D. (2001). *Essays on Actions and Events : Philosophical Essays Volume 1.* Oxford: Clarendon Press.

Davidson, D. (2005). Mental Events. In P. K. Moser, & J. D. Trout, *Contemporary Materialism* (pp. 111-126). London and New York: Routledge.

Deacon, T. (2013). *Incomplete nature: How mind emerged from matter.* New York: W. W. Norton & Company Inc.

Dennett, D. C. (1986). *Content and Consciousness.* Taylor & Francis Group. Retrieved from https://ebookcentral.proquest.com

Dennett, D. C. (1990). Cognitive Wheels: the Frame Problem of AI. In M. A. Boden, *The Philosophy of Artificial Intelligence* (pp. 147-170). New York: Oxford University Press.

DIMECC. (2017, March 3). *Design for Value*. Retrieved from DIMECC: https://www.dimecc.com/dimecc-services/d4v/

Dretske, F. (1988). *Explaining Behavior.* Cambridge: The MIT Press.

Dreyfus, H. L. (1972). *What Computers Can't Do: A Critique of Artificial Reason.* New York: Harper & Row.

Dreyfus, H. L. (2012). A History of First Step Fallacies. *Minds & Machines 22*, 87-99.

Dym, C. L., & Brown, D. C. (2012). *Engineering Design: Representation and Reasoning. Second edition.* New York: Cambridge University Press.

Endsley, M. R. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors 37(1)*, 32-64.

Endsley, M. R. (2015). Situation Awareness Misconceptions and Misunderstandings. *Journal of Cognitive Engineering and Decision Making Volume 9, Number 1*, 4-32.

Endsley, M. R. (2017). From Here to Autonomy: Lessons Learned From Human-Automation Research. *Human Factors*, 5-27.

Floridi, L. (2017). The Logic of Design as a Conceptual Logic. *Minds & Machines*, 495-519.

Fodor, J. (1974). Special Sciences (Or: The Disunity of Science as a Working Hypothesis). *Synthese, Vol 28, no 2*, 97-115.

Fodor, J. (1997). Special Sciences: Still Autonomous After All These Years. *Noûs, Vol 31, Supplement: Philosophical Perspectives, 11, Mind, Causation, and World*, 149-163.

Fodor, J. (2000). *Fodor, The Mind Doesn't Work That Way*. Retrieved from CogWeb: http://www.sscnet.ucla.edu/comm/steen/cogweb/Abstracts/Fodor_00.html

Fodor, J. A. (1985). Précis of The Modularity of Mind. *The Behavioral and Brain Sciences 8*, 1-42.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition 28 (1)*, 3-71.

Frankish, K., & Ramsey, W. M. (2012). *The Cambridge Handbook of Cognitive Science.* Cambridge University Press.

Franssen, M. (2013). Analytic Philosophy of Technology. In J. Olsen, S. Pedersen, & V. Hendricks, *A companion to the philosophy of technology* (pp. 184-188). West Sussex, UK: Blackwell Publishing Ltd.

Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems.* Boston: Houghton Mifflin Company.

Goertzel, B., & Pennachin, C. (2007). *Artificial General Intelligence.* Berlin Heidelberg: Springer-Verlag.

Harnad, S. (1990). The symbol grounding problem. *Physica D 42*, 335-346.

Heinämaa, S., & Tuomi, I. (1989). *Ajatuksia synnyttävät koneet.* Juva: Werner Söderström Osakeyhtiö.

Hempel, C. G. (1942). The Function of General Laws in History . *The Journal of Philosophy*, 35-48.

Hevner, A., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly, vol 28, no 1*, 75-105.

Insaurralde, C. C., & Lane, D. L. (2014). Metric assessment of autonomous capabilities in unmanned maritime vehicles. *Engineering Applications of Artificial Intelligence 30*, 41–48.

Jacob, P. (2014). *Intentionality*. Retrieved from The Stanford Encyclopedia of Philosophy (Winter 2014 Edition): https://plato.stanford.edu/archives/win2014/entries/intentionality/

Janssen, A., Klein, C., & Slors, M. (2017). What is a Cognitive Ontology, Anyway? *Philosophical Explorations Volume 20, 2017 - Issue 2: Cognitive Ontologies in Philosophy of Mind and Philosophy of Psychiatry* , 123-128.

Jaworski, W. (2016). *Structure and the Metaphysics of Mind: How Hylomorphism Solves the Mind-Body Problem.* Oxford University Press.

Jonas, H. (2001). *The Phenomenon of Life.* Evanston, Illinois: Northwestern University Press.

Kahnemann, D. (2011). *Thinking, fast and slow.* New York: Farrar, Straus and Giroux.

Kary, M., & Mahner, M. (2002). How Would You Know if You Synthesized a Thinking Thing? *Minds & Machines 12*, 61-86.

Kerr, P. (2014). Engineering Differences Between Natural, Social, and Artificial Kinds. In M. Franssen, P. Kroes, T. Reydon, & P. Vermaas, *Artefact Kinds: Ontology and the Human-Made World* (pp. 207-). Heidelberg New York Dordrecht London: Springer.

Kroes, P. (2013). Engineering Design. In J. K. Olsen, S. Pedersen, & V. Hendricks, *A Companion to the Philosophy of Technology* (pp. 112-117). Blackwell Publishing Ltd.

Krogmann, U. (1999). From Automation to Autonomy: Trends Towards Autonomous Combat Systems. *Advances in Vehicle Systems Concepts and Integration (RTO MP-44)*.

Kuhn, T. S. (1970). *The Structure of Scientific Revolutions.* Chicago & London: The University of Chicago Press.

Ladyman, J., Lambert, J., & Wiesner, K. (2013). What is a Complex System? *European Journal for Philosophy of Science*, 33-67.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences 40*.

Laplace, P. S. (1902). *A philosophical essay on probabilities.* New York: John Wiley & Sons. Retrieved from https://archive.org/details/philosophicaless00lapliala

Lawson, B. (1990). *How Designers Think 2nd Edition.* London: Butterworth Architecture.

Leppänen, M. (2005). *An Ontological Framework and a Methodical Skeleton for Method Engineering.* Jyväskylä: Jyväskylä University Printing House.

Levander, O. (2017). Autonomous ships on the high seas. *IEEE Spectrum (Volume: 54, Issue: 2)*, 26-31.

Lewis-Kraus, G. (14. December 2016). The Great A.I. Awakening. *The New York Times Magazine*. Noudettu osoitteesta

https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html

Lloyd's Register. (2016). *Cyber-Enabled Ships ShipRight Procedure – Autonomous Ships First Edition.* Lloyd's Register Group Limited.

Lützhöft, M., & Nyce, J. (2008). Integration Work on the Ship's Bridge. *Journal of Maritime Research Vol V. No. 2*, 59-74.

Marcus, G. (2018). Deep Learning: A Critical Appraisal. Noudettu osoitteesta https://arxiv.org/abs/1801.00631

Marr, D. C. (1990). Artificial Intelligence: A Personal View. Teoksessa M. A. Boden, *The Philosophy of Artificial Intelligence* (ss. 133-146). New York: Oxford University Press.

Mayr, E. (1974). Teleological and Teleonomic: A New Analysis. *Boston Studies in the Philosophy of Science Volume XIV*, 91-117.

Mayr, E. (2014). *What Evolution Is: From Theory to Fact (Science Masters).* Phoenix.

McCarthy, J. (2007). From here to human-level AI. *Artificial Intelligence 171*, 1174-1182.

McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. (1955). *A Proposal for the Darthmouth Summer Research Project on Artificial Intelligence.*

McCulloch, W. S., & Pitts, W. H. (1990). A Logical Calculus of the Ideas Immanent in Nervous Activity. In M. A. Boden, *The Philosophy of Artificial Intelligence* (pp. 22-39). New York: Oxford University Press.

McDermott, D. (2007). Level-headed. *Artificial Intelligence 171*, 1183-1186.

McLaughlin, B., & Bennett, K. (2018). *Supervenience*. Retrieved from The Stanford Encyclopedia of Philosophy (Spring 2018 Edition): https://plato.stanford.edu/archives/spr2018/entries/supervenience/

Metzinger, T. (2017). The Problem of Mental Action: Predictive Control without Sensory Sheets. In T. Metzinger, & W. Wiese, *Philosophy and Predictive Processing.* Frankfurt am Main: MIND Group.

Minorsky, N. (1922). Directional stability of automatically steered bodies. *Journal of the American Society of Naval Engineers*, 280-309.

Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In V. C. Müller, *Fundamental Issues in Artificial Intelligence* (pp. 553-571). Berlin: Springer.

National Science and Technology Council. (2016). *Preparing for the Future of Artificial Intelligence.* Washington, D.C: U.S. Government Office of Science and Technology Policy.

Newell, A., & Simon, H. A. (1961). Computer Simulation of Human Thinking. *Science, New Series, Vol. 134, No. 3495*, 2011-2017.

Newell, A.;& Simon, H. A. (1976). Computer Science as Empirical Inquiry: Symbols and Search. *Communications of the ACM*, 113-121.

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. *Computer Vision and Pattern Recognition (CVPR '15)*. IEEE.

Pahl, G., Beitz, W., Feldhusen, J., & Grote, K.-H. (2007). *Engineering design: A systematic approach (3rd edition).* London: Springer.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A Model for Types and Levels of Human Interaction with Automation. *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS – PART A: SYSTEMS AND HUMANS, VOL. 30, NO. 3*, 286-297.

Pearl, J. (2018). Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution. Noudettu osoitteesta https://arxiv.org/pdf/1801.04016.pdf

Penrose, R. (1990). Precis of the Emperor's New Mind: Concerning computers, minds, and the laws of physics. *Behavioral and Brain Sciences*, 643-705.

Peterson, J. B. (2013). Three forms of meaning and the management of complexity. In K. D. Markman, & T. (. Proulx, *The psychology of meaning* (pp. 17-48). Washington, DC: American Psychological Association.

Polvara, R., Sharma, S., Wan, J., Manning, A., & Sutton, R. (2017). Obstacle Avoidance Approaches for Autonomous Navigation of Unmanned Surface Vehicles. *The Journal of Navigation*.

Poole, D.;& Mackworth, A. (2010). *Artificial Intelligence: foundations of computational agents.* Cambridge University Press. Noudettu osoitteesta http://artint.info/html/ArtInt.html

Popper, K. J. (1979). *Objective Knowledge: an evolutionary approach.* Oxford New york: Oxford University Press.

Putnam, H. (1960). Mind and Machines. *Journal of Symbolic Logic*, 57-80.

Putnam, H. (1967). The nature of mental states. In W. H. Capitan, & D. Merrill, *Art, Mind and Religion* (pp. 1-223). Pittsburg University Press.

Revonsuo, A. (2001). Kognitiotieteen filosofiaa. In P. Saariluoma, M. Kamppinen, & A. (. Hautamäki, *Moderni Kognitiotiede* (pp. 51-84). Helsinki: Gaudeamus.

Ridel, R. (2015, October 2). *Mechanical Turing Machine in Wood*. Retrieved from YouTube: https://www.youtube.com/watch?v=vo8izCKHiF0

Roederer, J. G. (2003). On the Concept of Informatio and Its Role in Nature. *entropy 5*, 3-33.

Rohde, M., & Stewart, J. (2008). Ascriptional and 'genuine' autonomy. *BioSystems 91*, 424-433.

Rolls-Royce. (2016). *Remote and Autonomous Ships: the next steps.* Rolls-Royce plc.

Rosen, R. (1999). *Essays on Life Itself.* Columbia University Press. Noudettu osoitteesta https://ebookcentral.proquest.com

Saariluoma, P. (1992). *Taitavan ajattelun psykologia.* Helsinki: Otava.

Saariluoma, P. (1997). *Foundational analysis: Presuppositions in experimental psychology.* London and New York: Routledge.

Saariluoma, P. (1999). Neuroscientific Psychology and Mental Contents. *Lifelong Learning in Europe*, 34-39.

Saariluoma, P. (2009). From the Editor in Chief: The Conceptual Levels and Theory Languages of Interaction Design. *Human Technology*, 116-120.

Saariluoma, P. (2010). SCIENTIFIC AND DESIGN STANCES. *Human Technology*, 151-154.

Saariluoma, P. (2015). Four Challenges in Structuring Human-Autonomous Systems Interaction Design Processes. In A. P. Williams, & P. D. Scharre,

Autonomous Systems: Issues for Defence Policymakers (pp. 226-248). Norfolk, Virginia: NATO.

Saariluoma, P., & Rauterberg, M. (2015). Turing test does not work in theory but in practice. *Proceedings of the 2015 International Conference on Artificial Intelligence* (pp. 433-437). CSREA Press.

Saariluoma, P., & Rauterberg, M. (2016). Turing's Error-revised. *International Journal of Philosophy Study (IJPS), Volume 4*, 22-41.

Saariluoma, P., Cañas, J., & Leikas, J. (2016). *Designing for Life: A Human Perspective on Technology Development*. London: Palgrave Macmillan.

Saariluoma, P., Hautamäki, A., Väyrynen, S., Pärttö, M., & Kannisto, E. (2011). Microinnovations among the paradigms of innovation research - what are the common ground issues? *Global Journal of Computer Science and Technology, 11 (12)*, 12-24.

Schiaretti, M., Chen, L., & Negenborn, R. R. (2017a). Survey on Autonomous Surface Vessels: Part I - A New Detailed Definition of Autonomy Levels. *Computational Logistics. ICCL 2017. Lecture Notes in Computer Science, vol 10572.* (pp. 219-233). Springer, Cham.

Schiaretti, M., Chen, L., & Negenborn, R. R. (2017b). Survey on Autonomous Surface Vessels: Part II - Categorization of 60 Prototypes and Future Applications. *Computational Logistics. ICCL 2017. Lecture Notes in Computer Science, vol 10572.* (pp. 234-252). Springer.

Schierwagen, A. (2012). On reverse engineering in the cognitive and brain sciences. *Natural Computing*, 141-150.

Searle, J. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences 3 (3)*, 417-457.

Searle, J. (1984). *Minds, Brains, and Science.* London: British Broadcasting Corporation.

Searle, J. (1990a). Is the Brain's Mind a Computer Program? *Scientific American*, 26-31.

Searle, J. (1990b). Consciousness, explanatory inversion, and cognitive science. *Behavioral and Brain Sciences*, 585-642.

Shagrir, O. (2005). The Rise and Fall of Computational Functionalism. In Y. Ben-Menahem, *Hilary Putnam* (pp. 220-250). Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo: Cambridge University Press.

Shanker, S. (1998). *Wittgenstein's Remarks on the Foundations of AI.* London and New York: Routledge.

Shannon, C. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal, Vol 27*, 379-423, 623-656.

Shannon, C. E. (1950). Programming a Computer for Playing Chess. *Philosophical Magazine* .

Shoham, Y., Perrault, R., Brynjolfsson, E., & Clark, J. (2017). *Artificial Intelligence Index 2017 Annual Report.* One Hundred Year Study on AI at Stanford University.

Siegwart, R., Nourbakhsh, I. R., & Scaramuzza, D. (2011). *Introduction to Autonomous Mobile Robots (Second Edition).* Cambridge, Massachusetts: The MIT Press.

Simon, H. A. (1981). *The Sciences of the Artificial.* Cambridge, Massachusetts: The MIT Press.

Skolimowski, H. (1966). The Structure of Thinking in Technology. *Technology and Culture, Vol. 7, No. 3*, 371-383.

Taylor, C. (1965). *The Explanation of Behavior.* London: Routledge & Kegan Paul.

Thagard, P. (2005). *Mind: Introduction to Cognitive Science (second edition).* Cambridge & London: The MIT Press.

Thompson, E. (2010). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind.* Cambridge, Massachusetts: The Belknap Press of Harvard University Press.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, 433-450.

Veres, S. M., Molnar, L., Lincoln, N. K., & Morice, C. P. (2011). Autonomous vehicle control systems – a review of decision making. *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*.

Vernon, D. (2014). *Artificial Cognitive Systems: A Primer.* MIT Press. Retrieved from https://ebookcentral.proquest.com

Wiener, N. (1985). *Cybernetics: or control and communication in the animal and the machine (second edition).* Cambridge, MA: The MIT Press.

Wilden, A. (1987). *The Rules are No Game.* London: Routledge & Kegan Paul Ltd.

Williams, A. (2015). Defining Autonomy in Systems: Challenges and Solutions. In A. P. Williams, & P. D. Scharre, *Autonomous Systems: Issues for Defence Policymakers* (pp. 27-62). Norfolk: NATO Capability Engineering and Innovation Division.

Vincenti, W. G. (1990). *What Engineers Know and How They Know It: Analytical Studies from Aeronautical History.* Baltimore London: The Johns Hopkins University Press.

Winning, J., & Bechtel, W. (2016). Information-Theoretic Philosophy of Mind. In L. (. Floridi, *Routledge Handbook of Philosophy of Information* (pp. 347-360). London and New York: Routledge.

Winston, P. H. (20. April 2016). *12b: Deep Neural Nets*. (MIT OpenCourseWare) Noudettu osoitteesta YouTube: https://www.youtube.com/watch?v=VrMHA3yX_QI

Wittgenstein, L. (1922). *Tractatus Logico-Philosophicus.* London: Kegal Paul, Trench, Trubner & Co LTD.

Wittgenstein, L. (1994). *The Blue and Brown Books.* Oxford Cambridge: Blackwell.

Von Eckardt, B. (2012). The representational theory of mind. In K. Frankish, & W. M. Ramsay, *The Cambridge Handbook of Cognitive Science* (pp. 29-49). New York: Cambridge University Press.

von Wright, G. H. (2004). *Explanation and Understanding.* Cornell University Press.