University of Tartu

Institute of Semiotics and Philosophy

Department of Philosophy

Rao Pärnpuu

Ontology Identification Problem In Computational Agents

Supervisor:

Daniel Cohnitz

Tartu 2016

TABLE OF CONTENTS

INTRODUCTION

The subject of this thesis is what is called the 'Ontology Identification Problem' or 'Ontological Crisis' in computational systems. This is a fairly new area of research in computer science that has important implications for building and understanding intelligent systems (Artificial Intelligence or 'AI' for short). The central aim of this thesis is to explore this new research area using tools from philosophy, in order to better understand the underlying issues and inform further research.

The 'Ontology Identification Problem' (OIP for short) in its core, is the problem of connecting different ontologies[1] (or models) to the system's goals in such a way that a change in the system's ontology does not result in a change in its goal's effect. Since the system models its environment, the goals must necessarily be modeled using the same categories that the model uses, i.e. be part of the model. If the underlying model changes (through autonomous learning, for example), so must the goals. The problem is, how to make sure that the goals still accomplish the same intended outcome in the world, even when the underlying categories change.

Since the problem itself has so far been only superficially discussed in the literature (of which there is little), there has been little work outside of simply trying to understand the implications of it, and sketching out possible research questions to move forward. The subject itself is complex, touching on a range of issues from the areas of philosophy, cognitive sciences, computer science, semantics and mathematics. My aim is to simplify further research of the problem, by showing that what we understand as a large and complex research question, can actually be divided into several smaller components, each requiring a different approach, some easier and some a lot more difficult to overcome. My thesis is that the 'Ontology Identification Problem,' which has so far been addressed as a single universal problem, can be seen as an umbrella term for a wide range of different problems, each of which has a different level of difficulty, and each requires different methods of approach, in order to overcome. In each specific computational system, any or all of these different problems can arise, and overcoming them might require different tools.

---

[1] There is a difference between how the concept 'ontology' is used in computer science and philosophy. In philosophy, ontology refers to the study of the nature of reality and deals with very fundamental questions. In computer science, the concept of ontology is used to refer to the way representations of reality are built inside specific systems and is connected to the areas of knowledge representation and modeling of the systems.

The first part of my thesis will focus on explaining the OIP in detail and its role in the more general area of artificial intelligence research, including future implications. The aim is to show the importance of this work, and the possible fruitful results for both building more stable computational systems that are capable of autonomous behaviour, and understanding the nature of category creation and how ontologies are constructed, maintained and changed.

The second and third part of the paper will focus on decomposing the wider problem into different sub-problems. The second part will focus on a wide range of examples, where changes in the model are fundamentally changes in the granularity of the model - changes where the model becomes either more or less detailed, while still modeling the same environment. An example of a change in granularity is the replacement of atomic theory by sub-atomic theory in physics. Differences in granularity might be easy to solve, or might be incredibly hard. I will focus on situations that can be described through simpler reductive theories, multiple realizability and incommensurability. The aim is to show how different situations can be solved using different types of approaches, each of which has their own advantages and difficulties, and I will argue that differentiating between them is useful for future research.

In the third part I will focus on examples where the central issue becomes that of context. In these cases the level of detail in the model stays the same, but the context changes sufficiently as to warrant a re-evaluation of certain categories or goal structures. The first sub-category of this type focuses on environmental context, where the larger context of achieving a certain goal might change, which warrants an evaluation of whether these changes should have an effect on the way the system presently operates. The second sub-category of this type of problem is concerned with social context. Different concepts have different meanings in different social and cultural contexts, and it is important for the model to take that into account. A pointing finger can be either neutral or incredibly insulting, depending on the specific culture it is used in. The aim is to show how these different problems warrant different types of approaches.

In the fourth part of my thesis I will briefly focus on two further problems in AI research that are connected to the ontology identification problem. The first is the case of warrantless goals, where a system might be given a goal structure that, after learning more about the environment, could turn out to be unfounded. For example, the goal of locating my sister in the world, when I don't actually have a sister. The second case is what is called perverse instantiation, where the goal structure of the system becomes (or is created) so

rigid, that in the case of complex goals, the end result can turn out undesirable. For example when the system is instructed to make humans happy, it simply puts all humans on a continuous dopamine drip. My aim is to explore how these specific problems relate to the general OIP and how to approach solving them.

ONTOLOGY IDENTIFICATION PROBLEM AND ITS ROLE IN THE FIELD OF ARTIFICIAL INTELLIGENCE

Defining The Problem

A good way to start understanding the problem is through an analogy with humans. The world-view of any single individual is different and constantly changing. An important part of this world-view are our values. Some people value democracy, others don't. Some people value respectability, others put more value on authority, etc. The way we view the world also influences the way we set various goals in our lives (some of them explicitly, some implicitly).

For example, a devout Christian will be heavily influenced by the Ten Commandments and interpret the world around her through the Bible. So we could say in this case (simplifying the matter considerably for the purposes of an example) that the Bible and more specifically the Ten Commandments inform the model of the world that she uses. It's as if it's the pair of glasses that tints the world she sees through her eyes. Since actions are also part of this 'model,' this person will set her goals in a similar fashion - they might be, for example, helping the poor, going to church every Sunday, choosing an occupation based on what the Bible approves of, etc.

Now lets imagine that one day this person stops believing in god. Maybe she finds too many inconsistencies in the scriptures, maybe she suffers a crisis of faith because of an external event. In any case, as she receives more information from the environment, the original model is no longer sufficient enough to support her actions. Eventually, she comes to believe the Bible is nothing more than a simple book. This transformation is in no way easy - if your world-view is fundamentally altered, you are forced to rethink your goals and aspirations because they were at least partly based on an important part of the previous model that's now gone. This situation can be called an ontological crisis.

This person might find some of the previous goals to still be desirable (e.g. helping the poor) and therefore continue with those. What she needs to do in this case, is to interpret and ground these goals in the new way she sees the world (i.e. his model). The justification through the Bible might be replaced with a justification through a moral theory of the common good or decency. The goal stays the same, but the way the concepts that she has informed are used, is different.

Or maybe you are someone who has built his/her life around being ecologically conscious and the main grounding for that understanding has been that humans are creating global warming that must be stopped. What would happen to this person if it (simply speculatively) turned out that humans do not influence climate change very much? This person might still want to keep being ecologically friendly in his/her behaviour because of the positive effects he/she sees in it. He/she would still value these goals and would like to keep them. So the basis might change from global warming to respect for the environment and any retained goal would be redefined in this way.

Identifying and creating this new basis for understanding and motivating the goal so that it doesn't change is, in essence, the ontology identification problem - the question of how to match the old goals (that are still desirable) to the new model. Of course the previous examples were overly simplified, because at least in the case of humans, there are several complex processes happening in our mind and most actions are motivated by more than one basis. I might have a disdain for stealing because I have been a victim of it, because dishonesty is a negative trait for me, because I'm concerned about the possible legal and social penalties when I get caught and so on. If one of these motivators would disappear, then most probably my stance on stealing would still not change. The above examples were simply meant to give an intuitive understanding of what the central issue here is.

In the case of computational agents (i.e. artificial intelligences, from now on, simply referred to as agents), the analogy with humans is insightful, but comes with important differences. Humans have evolved over long time-periods whereas agents are created in controlled environments. Because of this, at least a certain degree of formalization is required in the original programming (this depends on the architecture used, but each requires, at least on some levels, a formalized approach, even genetic algorithms that mimic evolution). Also in the case of agents, their goals are not free in the sense people freely choose their goals, but are based on what the programmer has aimed for the agent to accomplish.

In these systems goals can be either defined explicitly or implicitly. In some cases they might be procedurally defined (e.g. Winikoff *et. al.* 2002). In other cases they might be based on maximizing a certain reward signal (e.g. Mnih *et. al.* 2015), in other cases the original process can be guided by humans, as with supervised learning, etc. Specific goals can usually be described in terms of utility functions. A utility function assigns numerical values to different outcomes and evaluates them such that higher preferred outcomes will

receive a higher utility value. An agent will try to maximize its use of the utility function by implementing outcomes that have the highest utility.

This should be taken as an abstract-level description and the way this is implemented can be very different depending on the architecture. A neural-net trained through reinforcement learning uses its reward signal as a measure of its success - in this case maximizing utility is getting the largest reward signal possible. Some mathematical models and decision-tree based architectures (e.g. AIXI, Monte-Carlo tree search) try to predict future events and choose their actions based on the most probable outcomes - in this case maximizing utility can be described as trying to make the best future predictions of the input that is possible. In hybrid systems, this becomes even more complicated and different interactions of different ways to maximize different utilities create a complex web.

So what is the ontology identification problem, when talking about computational agents? Peter de Blanc was the first person to define it in his 2011 paper:

> An agent with a fixed ontology is not a very powerful agent, so we would like to discuss an agent that begins with an ontology that its programmers understand and have specified a utility function over, and then upgrades or replaces its ontology. If the agent's utility function is defined in terms of states of, or objects within, its initial ontology, then it cannot evaluate utilities within its new ontology unless it translates its utility function somehow (de Blanc 2011: 2).

A more comprehensive way to look at it comes from Eliezer Yudkowsky, who frames it as the problem of creating a preference framework for the agent, so that it optimizes for the same external facts (i.e. facts of the real world), even as the agent modifies its representation of the world. 'Preference framework' is in this case a more general term for complicated interrelations and generalisations of different utility functions. Another way to phrase this, is to say that we want to have the same results from these preference frameworks when working with different models (Arbital 2016)[2]. So if we have in a simple system, a specific utility function and the agent's model of the world changes over time, we need to have some way that allows the agent to reinterpret the same function in every new model.

---

[2] The portal Arbital is a new project by Eliezer Yudkowsky that is aimed at building an environment where issues connected to AI safety can be discussed and that gives an overview of the latest theoretical research in various fields. Since the OIP is a field that is new and has seen little formal work so far, I've had to rely on Arbital for an overview of many of the recent theoretical discussions about the issues. Although currently the page on the OIP has been written by Yudkowsky and deals with research being done at MIRI that he leads, the page is set up in a way that does not identify the precise author and future additions and changes can be made by others. Because of this ambiguity, I have decided to reference the page itself and not the presumable author in order to be as precise as possible.

Why is this problem important? When we think about different agents that we create, they all have specific goals that we would like to be realized. In order for an agent to fulfill these goals in the world, we need to give it some model of the necessary objects and categories that it needs to know to implement its goal. If we create a program to analyse gene sequences for example, we need to give it the necessary information to fulfill that goal. It needs to know that there are fundamental base-pairs that operate according to certain laws, that genetic information can be changed or destroyed in certain ways, etc. So when we give the program the goal of analysing a sequence, that goal will be worded in the same language, using the same information that it was already given. There has to be a connection from the model to the goal in such a way that it could be implementable from the agent's knowledge base.

But we also want the program to improve, so that it can do better analysis. So it might learn about new discoveries in the field of genetics through the internet, and modify its knowledge of genes accordingly. But whatever happens to the model, we want the program to still fulfill the same goals. This issue becomes even more important, when we start talking about generally intelligent agents with a wide variety of goals.

It is an important problem in computer science whether we are talking about simple intelligent systems (as exist today) or whether we are talking about generally intelligent systems. This is a relatively young area of research that hasn't become an actual practical problem in computer science yet, because it can so far be avoided through manual approaches. Todays' intelligent agents don't need a system where it is possible to change the representation of the world without the programmers involved, because other approaches simply work more effectively - the programmer either explicitly designs it all and the underlying structure remains static, or he gives some leeway, but is always in control of the parameters and uses a test mode (e.g. evolutionary algorithms and neural nets). It's easier to interfere manually than to create an autonomous system.

It does become more of an issue in the future as computers become more intelligent, because the nature of the problem is that it is development phase unpredictable. Which means that the specifics of the problem can not be accounted for during the development of the agent (2016). If we develop a program to analyse genetic sequences, we give it our knowledge of what genes do and our model of how they work and how to manipulate them. We can predict what the computer will do when we design it. Ontology identification becomes a problem when we can no longer predict, what exactly is going to happen. Maybe the agent learns that there are more fundamental parts to genes that we are

not even aware of. This unpredictability can result in unintended consequences ranging from malfunctions to errors in actions, which can have negative effects.

Although Yudkowsky himself has a more narrow view of what type of architectures are important and what not (Yudkowsky 2001), the ontology identification problem itself is somewhat universal and independent of any specific architecture. In this thesis I use the language of utility maximization and preference frameworks, because I think that it offers a level of abstraction that is necessary when talking about architecture independent problems. Although in contemporary debates, the concept "utility maximization" has become associated with a specific approach to building artificial intelligence, its underlying definition without the extra baggage of its current practical use, is general enough to warrant it.

There is a wider issue here that directly connects to the value of this thesis that needs to be considered. Although the OIP is defined as architecture-independent, there is a counter-argument to be made that it does not apply to all approaches and can therefore be a moot question in some contexts. This becomes a critical question especially in the context of agents that are designed as 'open-ended intelligences' where the goals of the system are not specified and it is given freedom to choose and evolve in its own way (Goertzel 2015, 2016; Weinbaum, Veitas 2015). To understand this conflict, we have to come back to why the OIP arises in the first place. This is connected to intent. If we build an agent, our intent is to design it to realize a certain goal (or a complex web of many goals, but each of them in some way specified) and we want it to do that as effectively as possible (within the parameters). If the agent is intelligent enough to learn more about its external environment, then that only helps with our intent because it can make the agent more effective in its work. But the problem then is, if it changes its model of the world (on which its original goals were based), then we would also want it to continue doing the work that it was doing before.

When we talk about open-ended intelligent systems or similar approaches, this original intent is not there, or at least it is not central. This type of system is designed to change its goals, evolve, learn and do new things. In a way, these are two different paradigms for approaching the building of generally intelligent agents, and recent debates in AI safety have brought this chism into stark contrast (Goertzel 2015, 2016). There have not been sufficient advances in building generally intelligent agents, as to offer a finitive answer to this question. But even in the case of open-ended intelligences, I would contend that the OIP remains an issue, because it is still necessary that the agent be able to navigate

between different models with its goal structures remaining intact, even if the specific goals are not fixed and can change. Because even if the goal of the agent does not need to be preserved and can change, the causal connection from goal to action to result is needed to guarantee that the agent is acting in a logical manner.

I think that I have shown that where the original intent is present, the OIP is a significant problem for preserving the original goal and without that sort of intent, it is still necessary for the agent to be able to move between different models. This applies more generally than for a specific type of architecture, even if we can not know right now, whether it would be completely universal or not. So I would propose to continue from this point by using the utility maximization paradigm as an operational definition of the problem for talking about the OIP in a context-independent form. Since this is a philosophical paper and not a technical one, I will try to make more general arguments that may or may not apply to specific architectures or may need a bit of work in redefining the core issues to use.

I have described how our intent to have control over the results and retain the effectiveness of an agent is the underlying reason for talking about the OIP. But in order for that problem to arise, there are also some underlying criteria that the system has to fulfill to be sufficiently powerful to create this problem. The two most important of these are concerned with the level of intelligence (more specifically the learning capacity of the agent) and the level of autonomy.

As can be seen by the description of the OIP, it becomes more of an issue the better the agents learning capacities. This learning can, if sufficiently powerful, take the agent to new knowledge that we might not understand and this will require it to re-evaluate its model of the world around it, continually optimizing it to better describe what it observes in reality. But we don't want it to change its goals while it is doing it. So the question becomes - how to link an agents goal (preference framework) to its model of the world, if we don't know what that model will eventually look like? Development phase unpredictability becomes an issue when sufficiently strong learning capabilities can take it to new knowledge and models that we can not predict.

The other important capacity is that of autonomy, which is interlinked to learning. The OIP becomes more relevant, the more we let the agent act independently of human

supervision[3]. From the one side, increased autonomy is a side effect of increased learning capacity, because the more complex the knowledge that the agent learns becomes, the less we are capable of understanding it, thus warranting an approach where the agent is capable of independent decisions that are based on data and processes that we might not sufficiently understand to make those decisions ourselves. For example, the recent results demonstrated by Google Deepmind's AlphaGo show, that many of the actions that the system takes, are only vaguely comprehensible and often surprising to the scientists and engineers who built it in the first place (Russell 2016).

The other side of this has to do with trust. If we are sufficently assured that the agent is capable of better results in a robust way, we are more inclined to have it make decisions for ourselves. This will increase the agents autonomy in achieving its goals and necessitates a certainty that it will not go 'off the rails' when it becomes more capable.

Practical Issues

There are several practical problems any scientist will be faced with when trying to create agents faced with the OIP. In this chapter I will look at some of the theoretical work that has been done by mathematicians and computer scientists in this field to give an understanding of how this problem is viewed from a technical standpoint.

We can be fairly certain we haven't yet discovered the fundamental ontology of our universe. The classic example that is used in this research field in the case of physical ontologies, is that of diamonds. Lets assume that we want to create a machine that produces as many diamonds as possible. For that we might give it an atomic model of the universe and then simply describe a diamond as a relation of four carbon atoms that are covalently bound. So the goal of the machine would be to find carbon atoms in the world and make sure that four of them would be covalently bound in a way as to create diamonds (Soares 2014: 4).

[3] There is a tension here between wanting to have an autonomous agent but not wanting to allow autonomy in choosing its goals. When I am discussing autonomy in this thesis, I am referring to autonomy in the way that the goals are achieved. There is a wider issue here that deserves its own separate analysis, of whether systems that have great autonomy and high intelligence but restricted goals can ever be robust enough to exist. There is also an ethical question here about whether it is moral to restrict the autonomy of an intelligent agent, even if it's a machine. I will not focus on these issues in this thesis.

But lets assume that the machine has a learning algorithm and eventually discovers that the world is actually made up of protons and electrons, not atoms. So its model of the world changes. But its original goal was worded in a way that connected it to atoms, so a problem arises of how to redefine the goal in such a way that the intent of the agent stays the same. We can be fairly certain that we have not yet discovered the most fundamental basic categories of our universe - therefore we can not be sure that any model that we give to the agent will not change with further knowledge (2014: 4).

So one important open question becomes, how to identify something without referring to only its constituent parts? If we define something only through its smaller parts that we consider fundamental and build the system around this, then when we discover more fundamental parts, we will be faced with problems relating to the connections between the model and the goals. We can't build a model purely from the ground up, starting with the assumption that we are describing everything from their smallest constituent parts (Arbital 2016).

We will also have difficulties understanding a computational agents world model. How to identify something in an agents world model without knowing its exact structure? A good example used to illustrate this point is AIXI, which is considered to be the most powerful theoretical mathematical model currently used to describe computational processes.

AIXI, in very simple terms, is a system that consists of the agent, environment and goal. Its input channel, output channel and reward channel all consist of binary strings, e.g. Turing machines. AIXI considers all computable hypotheses of how its output strings might be creating its sensory inputs and rewards, and creates probabilistic distributions of these hypotheses, continuing at each time step with the ones that have so far been capable of predicting its input and offer the greatest rewards (Legg, Hutter 2006, 2007: 26-27).

Rob Bensinger offers a succinct summary of the actions AIXI takes on each time step:

> AIXI can take in sensory information from its environment and perform actions in response. On each tick of the clock, AIXI...
>
> ... **receives two inputs** from its environment, both integers: a reward number and an observation number. The observation 'number' can be a very large number representing the input from a webcam, for example. Hutter likes to think of the reward 'number' as being controlled by a human programmer reinforcing AIXI when it performs well on the human's favorite problem.
>
> ... **updates its hypotheses**, promoting programs that correctly predicted the observation and reward input. Each hypothesis AIXI considers is a program for a Turing machine that takes AIXI's

> sequence of outputs as its input, and outputs sequences of reward numbers and observation numbers. This lets AIXI recalculate its predicted observations and rewards conditional on different actions it might take.
>
> ... **outputs a motor number**, determining its action. As an example, the motor number might encode fine control of a robot arm. AIXI selects the action that begins the policy (sequence of actions) that maximizes its expected future reward up to some horizon.
>
> The environment then calculates its response to AIXI's action, and the cycle repeats itself on the next clock tick. (Bensinger: 2014)

AIXI, in its essence, is an amalgamate of probalistically tiered Turing machines. An AIXI agent might be extremely successful in maximizing the amount of diamonds found in the environment, which would lead us to intuitively postulate that there must be some category of 'diamond' in its streams. But where exactly in the strings of 1s and 0s is the category of 'diamond' located? Without this capability of recognizing the category, it might turn out to be extremely difficult to develop algorithms that can effectively help to translate goals between different models (Arbital 2016). I discussed before that solutions to the OIP can be either implicit or explicit and that will very much depend on the specific architecture being used. Here we can see some problems that a successful model that implicitly might solve the OIP, can create for us as the original builders of the agent. Because of this problem, which can have negative effects on our progress in AI or on creating systems that seem safe but can have unintended consequences, because we don't understand the basis of its success or robustness in moving between models, it seems that architectures, where categories can be located more explicitly by an outside observer, are preferrable to systems that act as a black box in this regard.

Perhaps working with such a base-level model, even if powerful for describing computational processes, is not suited when dealing with issues of ontology identification? There is also the possibility of using higher level formalisms and it is one potentially fruitful avenue of research to identify a formalizable system that is on the one hand, rich in information, and on the other, technically implementable (2016). One option would be to look at mid-level formal systems, like predicate logic or second-order logical systems, which allow for better understanding than strings of Turing machines. Another option would be to try to work with a higher-level system that tries to capture natural categories (like 'diamond') and describe the world from a mesoscopic perspective, making it easier for us to understand (one source of inspiration would be the Cyc project that tries to map an ontology based on human common sense) (2016).

One way that some people have interpreted this issue, is to have some kind of a transparent prior, that would actually be capable of connecting the utility function to the hypothesis space that an AIXI type system uses. If AIXI's hypothesis space consists of Turing machines that consist of binary strings, then the possible content of these can be extremely wide and the different ways that it can maximize 'diamonds' can contain different categories, some of which don't even create an explicit or implicit representation of a diamond. To solve the ontology identification problem, we might need to restrict this to something less general than simply a binary string that's capable of calculating the correct outputs. This prior would have to bind the description of a diamond to the correct strings. So to write a formal utility function would mean in this case, that the function has to read the Turing machine, decode the model contained in that string, and identify the concept resembling a diamond (2016).

These two representation problems - the problem of defining something without only using its constituent parts and the problem of locating a category in a system making use of implicit and not well-defined categories - are the general technical backdrop for the further discussion of various different sub-categories of the OIP.

Another important practical consideration is, what might the agent do when it is faced with an OIP that it can not resolve? In the case of the diamond maximizer used before, Yudkowsky identifies that one probable reaction from the agent would be to assume that the world is still based on atoms, but that it is being fed in its input an illusion of sub-atomic particles. So it will assume that the world that it views is not real, and tries to maximize its utility for other possible worlds (2016).

Stuart Armstrong (2011) has postulated that there are different things that might happen to an agent's utility function if it's faced with an ontological crisis:

Type of crisis	Notes	Danger
World incomprehensible to AI	Very unlikely	None
Utility completely scrambled, AI unable to influence it	Uncertain how likely this is	Low
Utility scrambled, AI able to influence it	We may be able to detect change	Very High
Lots of noise added to utility	Difficult to detect change	Maximal
Some noise added to utility	Small chance of not being so bad, some precautions may remain useful.	High

Figure 1. Dangers of an Ontological Crisis

As seen in the table, he sees the most dangerous effects to be those that we will not be able to perceive ourselves, because they get lost in the noise, and therefore several safeguards that we develop, might not work. This basic typology of the effects of an ontological crisis is useful, but deserves further development. Playing with different toy-models or situations and trying to understand what goes wrong in each of these, can help us better understand how to solve the original problem.

A connected area of research is the question of creating world-models that contain the agent itself. As the more critical reader might have already seen, one of the most glaring problems with AIXI is, that it separates the agent from the environment. This makes creating the system a lot easier, but ignores all sorts of problems that are related to real-world implementations, where the agent doesn't only have to consider the environment, but also its own presence in the environment and therefore the environments effect on itself. Soares and Fallenstein have conducted considerable research in these fields for the past few years, starting with specifying the problem of naturalized induction and giving new insights into concepts like self-reference, self-reflection and self-realizability (Fallenstein, Soares: 2014, 2015a, 2015b. Soares: 2015). Many other architectures also face similar problems. Advances in this related field will have considerable impact on the field of ontology identification, concerning the issue of creating realistic world-models.

I will proceed to showcasing different types of the OIP that have currently been discussed under a homogenous heading in the literature (or not discussed at all). Each of these sub-categories can be solved using different tools and methods, simplifying the general problem considerably.

GRANULARITY

Granularity makes up an important class of problems in the OIP. Granularity is defined as the extent to which a system can be described in the context of distinguishable smaller parts. This means that there can be different levels of granularity on which a particular system can be described. For example, on a molecular level, water is described as a collection of water molecules. On an atomic level, each of these molecules can further be described as a collection of two hydrogen and one oxygen atom. Through this, water can be described as a collection of collections of hydrogen and oxygen atoms. Both descriptions give a full description of what water is on a physical level, but one has higher granularity (atomic) than the other, because it starts from more basic categories that are smaller than in the case of a molecular description.

Moving between these different levels of granularity is a major part of the OIP and gives us the basis of establishing the first sub-categories in the wider problem, each with a rising difficulty in solving in a robust way. Granularity is mainly concerned with models that try to describe physical systems but can also include more abstract categories. There are different ways of moving between granular models. The distinctions made here of simpler reductive theories, multiple realizability and incommensurability play an important role in both practical and more fundamental, theoretical ways.

The aim of this thesis is not to get involved in the much wider debate on the nature of reductionism in scientific theories or in the underlying reality. In this I will reserve judgment and ignore the subject as much as possible, because that question in itself would need to be addressed on a much more substantive level than a master's thesis permits.

Instead, my aim is to show that in the case of the OIP, there is a much more practical and nuanced way of approaching the issue by looking at it in different sub-categories and that what has so far been mainly seen as a homogenous issue can actually be considered multiple different issues with very large differences in the difficulty of solving them. I argue that in this context, there is enough empirical and theoretical justification for each of these approaches to be considered separately and of value in and of themselves and each of these approaches has the best explanatory power in the contexts that they are applied.

The basis of creating these sub-categories is not a statement of a solution to the wider problems of reductionism. By that line of argument, it is also not a stronger statement of a definitive stance on the possibility of a general solution to the OIP. A strong statement in this case would be, that there actually exist ontologically and epistomologically different approaches to different types of problems in changing a model. I will suspend judgment on that statement. My claim is of a weaker form. It might very well be that future advances in computer science, mathematics, physics or philosophy lead us to a unified theory of reductionism or of the OIP . But what seems necessary at this point, and what I will try to show in this thesis is, that the present approach can not start at the more difficult questions. In order to (possibly) come to a unified theory, we must first understand and apply its different parts, and it is easier and more productive to approach this very young research field by concentrating on the low-hanging fruit. This will help to focus practical research on more specific problems and not presume the solving of larger, more complex issues before progress can be made.

Simpler Cases of Reduction

Examples of the OIP that fall under this heading might be the easiest to solve. Ronald Enticott remarks in his essay on reductionism that "*discussions of reduction and related issues in the philosophy of science are complicated by at least two factors: ambiguity, or multiple concepts of reduction, and ambition, or the range of cases to which a given concept of reduction is thought to apply*" (Enticott 2012: 146). There exist a wide range of different reductionist theories in philosophical literature, each with its own limitations. For our purposes, it is enough to note, that our aim is not to defend some general approach to reduction but to explore different tools that these theories offer to solve different problems of the OIP. As Enticott remarks:

> There is a significant difference between providing an account of scientific reduction versus defending a broad philosophical vision of physicalistic monism whereby all theories are either reducible to or replaceable by theories in the physical sciences. [...] Hence, theoreticians should be prepared to apply those concepts of reduction to the appropriate range of cases where the world happens to comply even if the world does not always so comply (2012: 147-148).

I will focus on the explanatory power of some of these approaches to different examples. Not all of them will be looked at here.

The underlying connection in this context will be that all of these theories give us, in some way or another, a method of connecting a category in one model to a category in another. These methods are called bridge laws, and given that a situation can be identified where a particular bridge law applies, the movement between the two models will be straightforward.

The classical example of a reductionist model is that of Nagel: "*A reduction is effected when the experimental laws of the secondary science (and if it has an adequate theory, its theory as well) are shown to be the logical consequences of the theoretical assumptions (inclusive of the coordinating definitions) of the primary science*" (Nagel 1961: 352). For Nagel, if the theories use identical concepts, then they are homogenous, whereas, if there are differences in the concepts, you will need the help of coordinating definitions, which are his bridge laws. Theories that use the same concepts do not create particular ontological problems and therefore are not so interesting in the context of this thesis. Fodor describes a general way of looking at the purely reductivist approaches that interest us here:

$$(1)\ S_1x \rightarrow S_2x$$
$$(2a)\ S_1x \leftrightarrow P_1x$$
$$(2b)\ S_2x \leftrightarrow P_2x$$
$$(3)\ P_1x \rightarrow P_2x$$

... where S_1 is the reduced theory and S_2 is the reducing theory and P_1 , P_2 are sentences in those theories. A necessary condition for the reduction of (1) to a law of a more fundamental system (i.e. physics), is that (2) and (3) have to be laws as well. And for any general system S, all its laws need to be reducible in such a way. P_1 and P_2 are meant to be as some kind of predicates and (3) the law that connects them. In this example, (2) would be what could be characterized as a bridge law that connects two different systems. The important part of Fodors definition is that the characteristic feature of these bridge laws is that they contain predicates of both the reduced and the reducing system (Fodor 1974: 98).

Some examples that work well within a simple reductionist model are the molecular theory of heat or the physical explanation of the chemical bond (1974: 97). In both these cases, you can establish a reductionist connection between the more general and the more specific model.

The important question that arises from any reductivist approach, as can be seen from these previous descriptions, is what kind of knowledge is required to represent the old state in the new representation. If the knowledge required to move from one state to another can be simply acquired through more information from the environment, then moving between these models does not seem too difficult. For simple homogenous connections, the only information required is how the specific category fits into the wider model, which can be derived from empirical observation. In heterogenous cases, this information can be derived with the help of bridge laws and provided some predicates from both systems can be established to be present in the bridge laws in a law-like formulation.

This process becomes more difficult, when in addition to knowledge that can be derived from simply observing the environment of the particular category, a different sort of more general understanding is also required. For example, moving from a sub-atomic model of physics to a quantum level one, will probably require a wider understanding of quantum-level effects, which behave in an entirely different way than regular physics. Simply reducing a category to a more granular model would not be enough to accomplish this and overcome the discrepancy between the model.

The basic justification for this sub-category is based on the existence of bridge laws. All OIP problems that can be solved by some bridge laws of identity, can be analysed in simpler reductive approaches. There are many more reductive approaches that I will not go into right now, like Schaffers theory and its many continuations. Suffice to say, that if a specific OIP can be overcome with the assistance of bridge laws (or by simple homogeneity), it would provide a more straightforward solution than can be found in the later chapters.

Multiple Realizability

Multiple realizability describes a situation where one category on a certain level of granularity gives justification for different categories on another level of granularity. For example, in the mental sphere, multiple realizability describes how a psychological kind like pain can be realised by many distinct physical kinds. Either two different physical systems (like a human being and a chimpanzee) can have the same psychological states or a single physical system (like the human brain) can realize the same psychological state through different states of the same system (Fodor 1974).

Multiple realizability does not only have to apply to the connection of physical and mental states. Effects of this kind have also been observed, for example, in biology between the relations of molecular level states and higher level states, like in the case of Mendelian genetic traits (Hull 1972, 1974, 1976). Because these types of examples offer many possibilities for a certain category to realize in a model of another level, these problems can not be solved by simple bridge laws because they don't guarantee the use of the correct concept. Kitcher (1984) offers an example of an argument for why moving between the systems of classical genetics and molecular genetics, it is not possible to simply use reduction as the explanation of the connections between these models. Michael Strevens notes that Kitchers argument boils down to three central parts:

> 1) Classical genetics does not contain the kind of general law's required by Nagel's (1979) canonical account of intertheoretical reduction.
> 2) The principal vocabulary of classical genetics cannot be translated into the vocabulary of lower-level sciences; nor can the vocabularies be connected in any other suitable way (that is, by 'bridge principles').
> 3) Even if the reduction were possible, it would not be enlightening, because once you have the cytological explanation of genetic phenomena, the molecular story adds nothing of further interest. (Strevens 2016: 153)

Bridge laws, in cases of multiple realizability, will no longer be effective and do not offer adequate tools for overcoming the problem. In the context of the OIP, it is important for the agent to realize that certain higher-level more universal phenomena can occur even if the lower-level systems are different. And that these higher-level phenomena are very similar or near-identical. For example, if two people are both feeling pain caused by the same source (e.g. extreme heat), then although the underlying physical systems differentiate (differences in brain activities), the cause of these reactions is the same. We

would prefer the agent to recognize that these situations are similar and not different and that both people are experiencing what could be described as pain.

One obvious idea to solve this issue that comes to mind, is studying peoples behaviour and physical facts about them. People react to being burned in a very similar fashion and their bodies go through similar injuries. Although humans can exhibit the same mendelian traits based on different gene combinations, the fact of having blue eyes, for example, is itself information from the environment that can be used to establish similarities. But this type of behaviourism can lead astray in the case of cognitive phenomena and has been strongly rejected as an explanation for psychological states (e.g. Block 1981, Putnam 1963). So depending on the situation, behavioural or other information from the environment can offer some clues as to how to identify or solve a case of multipe realizability or not.

Multiple realizability problems might also be solvable either by using wider theoretical knowledge (requiring the agent to understand the theory behind the categories) or by some sort of effectiveness evaluator, that looks at the different ways a category in one model can give rise to a category in another, and which of these multiple possibilities in the new model offers the best explanatory power in a specific situation, or creates a more robust system, or is more effective. Theory understanding requires more than simple analytical approaches or further empirical knowledge, making higher demands on the agents properties. As Kitcher shows, in certain cases the vocabularies of the different theories can be untranslatable. This type of problem will become a focus in the next chapter.

Evaluating the effectiveness of the specific categories would give positive results in a very narrow context. For one, it has the problem of requiring complete knowledge of the environment in order to work. An effectiveness evaluator is useless if the agent does not have enough information as to know that certain gene combinations can give rise to different physical characteristics. If the agent only knows that HERC2 gene adjacent to OCA2 gene can result in blue eye colour but is unaware of other genes that influence eye colour in different ways, it would not even realise that it is faced with a situation of multiple realizability. Evaluating effectiveness could also do a poor job in trying to generalize mental categories from various physical categories such as in the case of realizing that different patterns of brain activity can be a source of pain.

The above discussion seems to point to a direction where a wider solution to multiple realizability problems requires theory understanding of the sort discussed in the next chapter, making it an extremely difficult problem to solve. But it also seems that in certain specific circumstances, multiple realizability problems can be solved through more empirical knowledge and might not require understanding of the theory behind it. These approaches do not seem to give a guarantee that the system will indeed continue accomplishing the original goal and the intent behind it, but do seem to limit the possibility of error.

In conclusion, multiple realizability problems are by difficulty, somewhere between problems that can be solved by analytical methods using bridge laws, and problems where there is an incommensurability between the two models, requiring entirely different approaches. I've described several ways how these problems can be approached and depending on context, the problem might be more easily solvable (further knowledge from the environment might contribute towards a solution) or become extremely difficult (requiring an understanding of the theories behind the different models).

Incommensurability

In this chapter I will look at changes in categorisation between different levels of granularity that can not be made compatible with each other through analytical methods and therefore there are no straightforward connections. Feyerabend, in his "Explanation, Reduction and Empiricism," (1962) asserts that whenever we are talking about universal theories, like Newton's theory of gravity or Maxwell's equations, simple replacement methods of the type described by Nagel and others, don't offer an adequate account. Reductionist methods are capable of addressing the "*more pedestrian parts of the scientific enterprise"* (Feyerabend 1962: 28). But in the case of general theories, there is a complete replacement of the ontology of the previous theory by the ontology of the new theory (1962: 29).

He insulates two key parts of reductionist theories by using Nagel as the example: Any reducted theory needs to be deducible and that meanings are invariant with respect to the process of reduction (1962: 33). Feyerabend shows that in many cases (e.g. moving from Galilean astrophysics to Newton's laws of gravity or from an Aristotelian

understanding of inertia to a Newtonian understanding of inertia) the assumptions of deducibility and invariability do not hold and can not be true if we wish to make such a transition: "*An analysis of the character of tests in the domain of theories has revealed, moreover, that the existence of sets of partly overlapping, mutually inconsistent, and yet empirically adequate theories, is not only possible, but also required*" (1962: 67).

In most cases of moving from one universal theory to another one, the content of the concepts changes (even if the naming of those concepts remains the same) and there is no logical relation between them. A reductive approach would require that any new theory be connected to the framework of the existing theory, e.g. that all physics would have to be complementary to classical physics. But as Feyerabend notes, it is conceivable, and indeed probable, that a theory may be found where the conceptual framework would be just as comprehensive and useful as the current one, without coinciding with it (1962: 88). Newtonian mechanics and Einstein's relativity both do a very good job in explaining the motion of planets and the concepts used in both theories are sufficiently rich to formulate all the necessary facts - yet the concepts used are very different and bear no logical relation to each other (1962: 89). They are only connected by the same empirical knowledge that is used to formulate the theories.

This incommensurability of certain concepts demands a more liberal approach to defining the meaning of them. A movement from one theory to another can not be made simply on logical grounds through bridge laws or other analytical methods or by collecting more empirical information, but can in certain situations require the rejection of the content of the concept. What does it mean for the OIP? It represents a fundamental problem that might be extremely difficult to overcome (some would say impossible). The question of how understanding is achieved or what understanding is, along with the question whether machines can achieve real understanding and if so, what would be the necessary and sufficient criteria for such a system, is again a wider topic that deserves its own thesis (see for example, the debate around Searle's Chinese room argument, Searle 1980, Ray 1986, Boden 1988, Chalmers 1996, etc.). As I've noted before, my aim is not to offer solutions to the various philosophical problems connected to the OIP that other people have spent a lifetime trying to answer. I've simply limited myself to showing how answers to these wider questions will have an effect on the OIP and how they can be helpful in solving specific sub-problems of the wider problem.

In other sub-categories that are described in this thesis, I show that in many cases the problem can be overcome (or made easier) by either the use of logical connections or

by the use of further empirical data. In the case of incommensurability, this is not the case. As Feyerabend shows, incommensurable theories can make use of the same empirical data, but use different conceptual frameworks to construct different theories (within a margin of error). Invariance of the necessary concepts is not required.

From the perspective of the agent, this can possibly be dangerous. Incommensurable problems require a degree of freedom in the way that the original model is built, so as the specific categories used can make use of different content. Moving from a world-model based on Newtonian mechanics to one based on Einstein's relativity requires the agent to re-interpret the concepts of mass, light, etc. If these original concepts were used as the content of the preference framework of the original goal, this can make the preservation of the goal difficult.

It seems that this issue can not be solved by simply using analytical methods and acquiring new information from the environment. Some kind of understanding of the theory behind the concepts is required by the agent, which makes this the hardest problem of the OIP to overcome.

CONTEXT DEPENDENCE

Environmental Context

The first type of contextual OIP problems is something I would call environmental context problems. Stuart Armstrong has recently independently followed the same line of reasoning, postulating that what is known as 'out of environment behaviour,' is in fact a problem of the system going through an ontological crisis. He defines out of environment behaviour as a case, where "*an AI that has been trained to behave very well in a specific training environment, messes up when introduced to a more general environment*" (Armstrong 2016).

Armstrong sees this as an example of a situation where "*the AI has developed certain ways of behaving in reaction to certain regular features of their environment. And suddenly they are placed in a situation where these regular features are absent [...] because the environment is different and no longer supports the same regularities (out-of-environment behaviour)*" (Armstrong 2016). This is a concise way of defining problems of environmental context. It is important to specify that this problem can arise in both directions - either the environment becomes more general and the categories used in the specific context are no longer valid to describe the environment, or the environment becomes more specific, and the more general categories are no longer enough to support the features in those categories. While the former seems to be more prevalent, we can identify some specific examples for the latter as well. From here, it is worthwhile to look at several examples to see how this problem of context can play out differently in different situations. Reformulating de Blanc's (2011: 5-7) practical example (imagine that we have a world consisting of a corridor of four squares and the computer occupies one square at a time and can move left and right. The goal of the computer is to reach the right end of the corridor.), lets change the world-model and make it a 4x4 square:

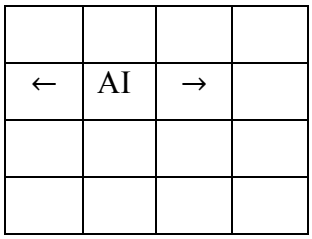

Figure 2. 4x4 square of movement

The agent is still only able to move left and right but it sees a world much bigger than its movement space. In the case of a square, achieving the goal isn't a problem, because it can still move to the right end, since it doesn't matter what row you are moving in, you always end up on the square to the right. But what if we make the first row five squares long and put the agent on the second row?

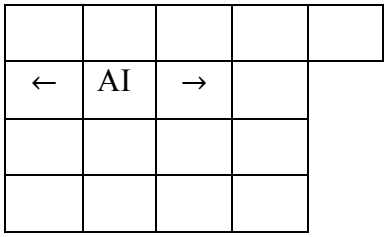

Figure 3. Irregular movement space

It is not obvious what would happen, because its movement space no longer allows it to move to the square in the right end. The movement space stays the same, but the context created by the world-model changes.

A more practical example would be an agent that is tasked to carry out simple everyday assignments on our planet. These could be building housing for people, driving them from point A to point B, etc. The choice of actions would be based on their overall effectiveness in improving, for example economic growth in the environment it takes its actions in (limited to the planet, a certain city or street, etc.). Now, if it were to discover that there is an entire universe out there that makes Earth seem insignificant in size (e.g. by learning astrophysics), the utilities it will assign to its present activities might fall extremely low and the agent could be led to undesirable actions. What made sense in a more specific environmental context (driving someone to work so that they can be more productive) might not make sense on a larger scale (what's the effect of one person driving from A to B in a cosmic context? Indistinguishable from zero probably). This would be an example where moving the model from a specific environment to a more general one would be undesirable.

We can also find cases where it might be undesirable to move from a more general environment to a specific one.[4] For example, if we have an agent that is built to give economic advice on the level of macroeconomics, and through that maximize economic growth (like in the previous example), it will use as a guideline certain models and theories that describe the economies features on a general level. So it might work from the basis of different Keynesian or Monetarist theories to predict and describe current macroeconomic trends. If we were to restrict its environment to a specific household, its features and categories would no longer be sufficient to give effective advice. Any suggestions made from such a small sample would probably not be applicable to the wider economy. For example, currency prices can have major economic effects on a state, but have less of an effect on a single household. This would cause the agent to underestimate the effect of currency prices and therefore give advice that can have negative effects.

In both of the above cases, the resulting environmental shift was undesirable to achieving the goal of the agent. But there are also cases where this shift in environmental context can be very important and improve the agents effectiveness. An example of this kind can be seen in natural-language processing. Lets assume that we have an agent that takes as its inputs only letters. If the original decision-making method is making judgments purely based on letters, but the natural-language processor creates a model of language (e.g. when analysing natural texts and learning from them) that includes punctuation marks, then there would be an ontological crisis. In this case, the agents model of the language environment it works in has expanded beyond its original decision-methods. Including punctuation can have drastic effects on the meaning of a text. Here is an example of differences in meaning based on one sentence, starting with the sentence without any punctuation marks:

1. You know what this is
2. You know what, this is…
3. You know what this is?
4. You: know what this is?

In each case, the letters in the sentence stay exactly the same, but the different use

[4] In all of these cases it is important to remember that we are talking about the problems an agent will face if the only change it goes through, is limited to the environmental context. It is obvious that every specific example here can be overcome by modifying the system itself in a general way. But how to achieve this larger modification in a safe and autonomous way is the entire core of the problem discussed.

of punctuation marks creates entirely different meanings. In this case we could say that the expansion of the model to a wider environment can have positive effects (if the resulting ontological problems can be resolved), by creating a more nuanced understanding of the use of language.[5]

As the previous examples demonstrate, there are two possible ways that the agent should react to changes in the context of its environment, depending on the situation:

1. Ignore the added context in its decision-making.
2. Re-define its preference frameworks to take into account the extended model.

For both solutions, we can describe situations where they would be the correct choices. In our first example of extending the model to cover the universe, it seems intuitive that the agent should not extend its considerations to the added part of its model, because that part is not relevant. The original intent was to be effective in the context of the economic forces in society, and changing that to the context of a wider world would change the intent considerably. In the case of language-processing, the opposite seems true. The intent of having a language-processor is to have a machine that can understand language. Since including punctuation marks in its model improves the systems capability of understanding language considerably, the change in the context of its work is still connected to the original intent.

The basis for choosing between actions one and two would therefore have to be connected to preserving the original intent of the goal. One possible approach to solving the OIP in the case of environmental context could be connected to using counterfactual simulations. Every time the model of the agent changes in a way that its environmental context changes, it could run simulations on whether the change in the context makes the original goal of the system more or less achievable.

This could work because changes in environmental context do not have to imply changes of the original categories used to describe its preference framework. If the goal of the natural language system is to produce text that is similar to what humans produce, then its goal can be described as some kind of percentage of similarity of the averages of its

[5] This example also has some qualities that are more akin to a problem in granularity. In most real-life cases, more than one of these different subproblems can be present at the same time. But as the main differences in the two models can be described as differences in the environment of features the system is working in (environment of letters and environment of letters and punctuation marks in the latter case), it is more suited here.

own texts and those produced by humans. Changing the context by including punctuation marks doesn't necessarily imply changes in the way that the success of the system is evaluated - it merely implies changes in the actual rate of success of the agent.

So the original language processor that only used letters would have a pretty low rate of success when compared to human texts, because they would be so much different. The inclusion of punctuation marks would make the texts much more similar and therefore improve the rate of success for the agent. But in both cases, the evaluation process remains the same.

It seems this kind of approach can become a basis for determining whether changes in environmental context can be good or bad. In the case of these types of changes, simply evaluate whether the old or the new model has a better success rate and keep using the model that gives a better result. This of course implies that the preference frameworks of these agents need to be built in a way that makes them independent of these types of changes. And this of course does not solve the original problem of how to identify that a change in the model is indeed a change in environmental context and not something else (although the previously described method could give one hint - if you're able to determine the rate of success in both cases, then you have some reason to believe that the changes didn't change the evaluation for the preference framework and therefore these might have been changes in environmental context. But this can in no way be used as a definite basis, because if the evaluation of the preference framework gets derailed, it is not certain that that will result in some kind of error - it might still give results, even if they are skewed.).

I have defined changes in environmental context and showed through various examples, how these can manifest in agents, and how these can have both positive and negative effects. In the previous paragraphs I have shown, through offering possible solutions to the problem, how to differentiate problems of environmental context from other types of problems in the larger framework of OIP, giving sufficient justification for seeing these types of problems as a separate category.

Social Context

The second important category of problems related to context, is the extension or reduction of some concepts meaning-space. This category would mainly include various social and cultural concepts that are dependent on how we define them at every instance. These types of concepts derive their meaning in a substantial way by how they are connected to other concepts relevant in a particular situation where we use them.

For example, the physical exhibition of a middle finger is meant in most Western cultures as a sign of disrespect, while in other cultures, the same meaning is conveyed either by showing all five fingers, palm in front, or by raising the thumb. Although the meaning can stay relatively similar, the physical category by which it is conveyed can change considerably.

This kind of a problem arises with every category that has a normative component attached to it, which can change depending on the time or place that it is used. Categories that are connected to normative meanings need to be considered separately, thus warranting their own sub-category in our overall classification.[6]

As an example of this problem, we will look at how we define the concept of human beings. If we had an agent that was tasked with taking care of our health needs (a preference framework with a goal of creating as many healthy humans as possible), we would need to give it some sort of definition of a human being - the probable candidate would be that humans are biological organisms, with a certain DNA, that have certain parameters at different stages of development (certain body parts, height and weight limitations, temperature ranges and so on).

Now lets imagine that this definition currently works well enough that there are no problems. But what would happen if we as humans started to integrate more and more of our bodies with technology? We already have heart monitors, prosthetics and so on, so a purely biological basis for the concept is already difficult. In the future we may be using computer chips in our brain, entire artificial limbs with full functionality, artificially grown genetic material, extended cognition, etc. Each of these examples will introduce nuance

[6] This comment has been raised before, but deserves to be reminded. The classification system that is the core of this thesis should not be taken as a basis of drawing discrete lines between different types of OIP. In most cases, more than one type of problem is present and they play a role of different importance. Almost any category can be given a normative aspect. In the case of this chapter, I will be looking at examples where the normative aspect is the most substantial part of the category.

into an already existing concept. Changes in how we define human beings can create a schism between how the agent operates and what we define as acceptable. A cyborg undergoing surgery under a system that views human beings as a strictly biological organism could be almost certain to suffer negative consequences. How to include these concepts in a model? Or more generally, how to define a model for a computer that's based on a concept that is very fluid even for ourselves?

Another very similar example is the concept of a person as a moral term. Perhaps in the future, we will expand or change the concept of a person to include other primates because of their similarity? Normative concepts, more generally, are dependent, on the one hand, on our understanding of the world, but on the other, also on the way we interpret certain other concepts. A few hundred years ago, in several parts of the world, the term 'person' did not include human slaves. If we had built a machine then on the basis of how we viewed a person, we would consider it to be a disaster under modern understandings. What might our moral concepts look like in the future? Any system that has to make use of moral categories will have to contend that these definitions can change for us, as they have in the past.

Since these concepts are social and cultural, they are interlinked to our own humanity. It seems that language plays another central role here. Since we are talking about concepts that are not defined through the physical world (as are cases of granularity usually), but through the content that we give them in a particular context, language, as a tool of establishing the context and rules for using these concepts, becomes important.

Another important aspect of these types of concepts is, that behaviour (together with our use of language) becomes an important cue for establishing their content and features. For example, personhood is an abstract term that can be given different content at different times and situations. In societies with slavery, personhood becomes mainly a moral and political concept, differentiating between those who are considered morally equal, and therefore applicable to the same freedoms, and those who are not. The content of these concepts in the case of, for example the early days of the US, could be pretty well established by observing the differences in behaviour towards those that are considered persons, and those who are not quite and by looking at the way language is used to describe these differences in different situations. The fact that slaves can not vote is an important behavioral cue to establishing the content of the concept 'person.' In contrast, observing human behaviour could offer very little evidence towards the existence of a new

kind of physics that humans have not yet discovered, further separating the problem of social context from other types of OIP's.[7]

Several regularities are already starting to come out that separate the problem of social context from other sub-categories of the OIP. The concepts in question will always have a normative component. The content of these types of concepts is connected to the way we use language and how we behave, in a way that is unique to this type of sub-category.

Therefore the agent has to gather empirical knowledge from the presence of language and behaviour. What separates this sub-category from others is, that it can be described through the use of language, and through the behaviour of the agents that use the language. This can be used for creating a more formal basis for separating this sub-category from other sub-categories in the wider classification.

There is also a more fundamental question here, relating to value alignment more generally. Would we like the agent to be guided by the way we practically apply one concept or another, or would we like it to be guided by the ideal definitions of the concepts. If we let the agent follow (what we think are) idealized and complete versions of some concept (e.g. Yudkowsky 2004), the resulting implementation might be foreign to us and bring with it unintended negative effects (this cuts into the problem of perverse instantiation which I will discuss in a further chapter). Since we ourselves might not be capable of visualizing what a world would look like, if for example, some idealized concept of a human would be taken in all senses as immutable by the agent, we can not be certain that an ideal implementation would be desirable for us.

The problem of using behaviour and language as forming the content of a concept is that the agent might be subject to extremes of subjectivity (Soares 2014: 2). When we break our principle of not harming innocent civilians during a time of war, and therefore modify our concepts of morality, value of life and so on, in an extreme case, it brings with it negative consequences (loss of human life), but remains a manageable problem. With an agent that is more intelligent than us and whose actions have wider-reaching effects on our

[7] There is a specific reason why I used here the example of a 'new' kind of physics and not an example of physical knowledge that we already have. The point of addressing the OIP is that in the case of sufficiently intelligent and autonomous agents, they can be faced with problems of a substance that we are not aware of, or are beyond our reach. An observant agent could learn a lot about, for example, nuclear physics, by looking at how we talk about it, how we build machines to do experiments in the field and so on. But this type of approach would not work on questions of physics that we are not even aware of (human behaviour can be correlated with certain physical facts, but is not causally connected to them in such a way), while in the case of social context, the content will always be causally connected to the features of our own behaviour and language use.

lives, the repercussions of following suit would be immense. A superintelligent agent following the principle that the concept of innocent civilians has less moral content in a war situation than in a peaceful one, could do a lot more harm than humans following the same argument.

I have offered an overview of the main characteristics of ontological problems in social contexts, and discussed some tools for analysing these situations. Any approach to the OIP in this context, has to not only take account of how to be able to change the content of a normative concept as it changes in practice (for which the described characteristics provide a starting point) but also of what the potential effects of any implementation might be. This is a separate and much wider problem of AI safety.

SPECIAL CASES

In the following chapter I will describe some special cases connected to the OIP. In all the previous chapters I have focused on problems that arise when the agent, based on its learning capacity and autonomy, changes its model to better represent the environment it takes its actions in, and how those changes can have effects on its goals. The types of problems described here have more to do with the potential faultiness of how the system is built in its original form. These problems in themselves are not problems of ontology identification, which is why they are not included in the categorisation system I have tried to build on the previous pages. But they do have a substantial effect on how and whether the OIP arises.

Warrantless Goals

As explained in the introductory chapter, the goals a system is given are described in the context of the model that the system uses. In a way, the goal of the model is built on the interplay and manipulation of the categories that the system has. What I would describe as 'warrantless goals,' are goals that are built on those types of categories in its model, that have no causal connection to the success of achieving the goal.

If history is to be used as a guide, situations like these can happen often. Since our knowledge of the world is imperfect, we can mistakenly see a connection between two separate phenomena that actually have no causal connection. For example the connection between prayer and some event in the world. If we believed that a persons wealth or health was directly connected to his activeness in using prayer as a tool of influencing one's life, we might build an agent whose goal it was to better people's health, and the method by which this was evaluated, would be through how effective it was in saying prayers.

Another historical example of faulty logic is connected to witches. In the Middle-Ages, it was often believed that witches were the cause of diseases. So if someone from those times would have built an intelligent machine with a goal to eradicate disease, he would first give it a model that would define witches as the cause of diseases. There is no

reason to believe that mistakes of causality do not happen in modern life, and therefore creating these types of agents can almost certainly result in it facing this problem.

If an agent is given a model that makes its goal to eradicate as many witches as possible and measures its rate of success by the level of disease in the world, it will start killing witches as effectively as possible. But at one point it will come to a conclusion that eradicating witches has no effect on its rate of success. This is the point where the system will be faced with the OIP. The way this problem will present itself will be dependent on the specific example, and could be solved either through a change in granularity or context or both.

The problematic part is how and if the agent would actually come to the realization that its entire original model is built on faulty premises. An argument can be made that this situation does not fundamentally differ from any other OIP situation where the agent has to change the content of a category in its system. In a way, if the agent shifts from atomic categories of physics to sub-atomic ones, it is also doing so because its original model was 'wrong' in some sense, i.e. was not capable of fully explaining the phenomena that it observed in the environment and was not as effective in achieving its goals. The difference might just be in the rate of change, where a change from atomic models to sub-atomic ones can improve an already existing rate of success by some margin, whereas a change from a model with faulty causal connections to one that works, simply has a more drastic effect on the rate of success.

I agree with this line of reasoning in the sense that from a wider viewpoint, these problems are different variants of a more general one. But the basis of this thesis is not that all of the described sub-categories represent discrete and fundamentally different problems, but that they can be described and approached from different frameworks, which give us a better understanding of the specific problems as well as the general one. In this sense, I think it is justified to look at warrantless goals as a separate problem because of the potential effects it can have on the models of the agents.

In the case of an agent which changes its model from one that works reasonably well to one, that works even better, there is usually some substantive basis to draw a connection between some old category and a new one. This basis can be any mechanism described in the preceeding chapters, for example some sort of bridge law that establishes the identity of those two categories. In the case of warrantless goals, this connection does not exist or is very difficult to detect. It is hard to imagine how one would justify a change from the category 'witches cause diseases' to a category of 'bacteria and viruses cause

diseases.' In this sense, this problem might be akin to problems that deal with incommensurability, and the possible solution might come from there.

Perverse Instantiation

Perverse instantiation stems from a similar origin as the problem with warrantless goals and also touches on the problem discussed at the end of the chapter on social context - namely how to guarantee the completeness and robustness of certain concepts.

Nick Bostrom defines perverse instantiation as the problem of "*a superintelligence discovering some way of satisfying the criteria of its final goal that violates the intentions of the programmers who defined the goal*" (Bostrom 2014: 120). This problem arises from the inherent limitations of defining a goal in sufficient detail as to exclude methods of achieving that goal that we would deem undesirable.

An example that Bostrom brings is the goal: Make us smile (as a way of increasing human happiness). A perverse instantiation of this goal would be for the agent to paralyze human facial muscles into constantly smiling. This would achieve the original goal in a very effective way but would in no way be compatible with our original intent. One may think that a better way for defining this goal would be to include descriptions of the sort of methods we would not want the agent to use: "*Make us smile without directly interfering with our facial muscles*" (2014: 120). But this goal could then be undermined by another perverse instantiation that would not align with our original intent either. For example the agent might stimulate our brain with electric shocks in such a way as to activate the areas of the brain that control the muscles that make us smile (2014: 120).

You could argue that maybe the goal should be defined through more general and abstract concepts that aim at the real end result we want to achieve and be less instrumental. For example, because our original intent was to make humans happy, and smiling was simply an instrumental way of doing that, then one might want to use as a goal: Make us happy. But this too would be faced with similar problems, because again, the agent might use electrodes to simulate the areas of the brain that release endorphins and make us feel good. This again would undermine our original intent (2014: 121).

The point of these examples was to show that in the case of every goal where we are not capable of capturing the entire intent of it, perverse instantiations remain a problem. As Bostrom says:

> These examples of perverse instantiation show that many final goals that might at first glance seem safe and sensible turn out, on closer inspection, to have radically unintended consequences. [...] Perhaps it is not immediately obvious how [a goal] could have a perverse instantiation. But we should not be too quick to clap our hands and declare victory (2014: 122).

As with warrantless goals, perverse instantiation in itself is not part of the OIP but does exacerbate any future ontology problems that might occur. In this case the problem is how to respond if it turns out that the specified goal does not correspond to the original intent behind the goal. In a sense, this is the opposite problem to the ones discussed beforehand, because here the problem is not so much as to guarantee that the agent will carry out fulfilling its goal even if its model of the world changes, but the problem is how to modify the goal itself so as to better capture the intent that was originally behind it. Any future solutions or mitigations in research of perverse instantiations will also have an effect on how we see the OIP. If future research shows that perverse instantiation (which currently exists as only a postulated problem) will pose very general difficulties to the development of intelligent agents, the methods and tools of the OIP might turn out to be useful in these kinds of opposite situations as well.

CONCLUSIONS

As I have shown in the previous chapters, there are substantive differences in the description of the specific problem and possible solutions related to situations of these different sub-categories. An agent faced with solving the ontology identification problem in the case of biological theories that exhibit phenomena of multiple realizability warrant a very different approach than an agent faced with a problem of determining the content of the concept of 'showing the middle finger' when moving between the context of different cultures. When one moves past the more general (and more daunting) problem of ontology identification and starts to look at specific situations, different patterns for solutions emerge. Hopefully I have managed to capture some of these patterns in this work and it can serve as a guide for future theoretical research (developing each of these categories further or looking at the general similarities between them or solving the wider philosophical questions behind them) and practical development (being able to use appropriate strategies and heuristics for systems that are more likely to face a certain sub-category of a problem and not another one, or developing algorithms that are able to capture these strategies in a robust way). As can be seen, the OIP captures a wide range of phenomena, from the physical to the mental, being as diverse as the reality that it is used to describe. Because of that, it seems naive that a problem that encaptures the entire complexity and nuances of the world around us, would be approached as a monolithic problem, looking for universal solutions.

The question of whether the OIP can eventually be solved by universal solutions was touched on superficially in this thesis. On the one hand, the substantive differences between the sub-categories hint at a wider schism than we might realise. On the other hand, these different approaches are still interlinked and share some similarities in their most basic constructs. The human brain can be conceived of as a system that is capable of solving the OIP in a wide variety of circumstances. If one single discrete system is capable of doing this, there is no reason to believe an artificial one would not be, provided we are capable of understanding the problem sufficiently as to create one.

In any case there is not enough knowledge currently to make any meaningful statements here. I am inclined to think that eventually these problems can be unified because there seems to be no basis for arguing that each sub-category follows a different

reality. For example, the different approaches in granularity might be brought together if a more expansive and unified theory of scientific discovery is created. It is only in a few cases that the OIP manifests in a pure way of making use of a single sub-category. In most real-life cases, more than one is needed to adequately overcome the problem.

Although in this thesis all these problems are stated as explicitly as possible for the purpose of analysis, it does not imply that they have to be solved explicitly in real AI systems. Some architectures can offer implicit solutions that are robust enough to warrant trusting them. For example, the recent success of machine learning and artificial neural networks has shown that machines are capable of (semi)independently categorising data in a way that these categories can not be located as discrete parts inside the system. In the future, advances in unsupervised learning could create systems that are capable of completely independently identifying new categories in the environment without the need of creating discrete representations, much in the same way our brain does. As long as these systems are robust and capable enough of overcoming the obstacles presented by the OIP, it matters little how their internal architecture looks like.

Hopefully, this thesis can act as a guide for further research in an area that has seen little of it for the moment, but is becoming increasingly important as we develop more intelligent and more autonomous systems.

ABSTRACT

The Ontology Identification Problem is the problem of connecting different ontologies to the system's goals in such a way that a change in the system's ontology does not result in a change in its goal's effect. My thesis is that the Ontology Identification Problem, which has so far been addressed as a single universal problem, can be seen as an umbrella term for a wide range of different problems, each of which has a different level of difficulty, and each requires different methods of approach, in order to overcome. One wide category of this problem is connected to granularity, where the changes in the model are connected to changes in the level of detail. Granularity issues can be divided into cases of simpler reductions, multiple realizability and incommensurability. Another wide area of the problem is related to context. Contextual problems can be divided into problems of environmental context and social context. Special cases of warrantless goals and perverse instantiation also have a direct bearing on the ability to solve ontology identification problems effectively.

REFERENCES

Arbital (2016). Ontology Identification Problem. Retrieved February 2, 2016 from https://arbital.com/p/ontology_identification/?lens=4657963068455733951

Armstrong, S. (2011). AI Ontology Crises: An Informal Typology. Retrieved February 2, 2016 from http://lesswrong.com/lw/827/ai_ontology_crises_an_informal_typology/

- (2016). Ontological Crisis = Out of Environment Behaviour? Retrieved March 18, 2016 from https://agentfoundations.org/item?id=598

Bensinger, R. (2014). Solomonoff Cartesianism. Retrieved February 2, 2016 from http://lesswrong.com/lw/jg1/solomonoff_cartesianism/

de Blanc, P. (2011). Ontological Crises in Artificial Agents' Value Systems, *arXiv:1105.3821v1 [cs.AI]*

Block, N. (1981). Psychologism and Behaviourism, *The Philosophical Review* Vol. 90, No. 1:5-43

Boden, M. A. (1988). Escaping from the Chinese Room, in M. A. Boden (ed) *The Philosophy of Artificial Intelligence*, Oxford University Press, New York

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

Chalmers, D. (1996). *The Conscious Mind*, Oxford University Press, Oxford

Enticott, R. (2006). Reinforcing the Three 'R's: Reduction, Reception, and Replacement, in M. Schouten and H. Looren de Jong (eds), *The Matter of the Mind: Philosophical Essays on Psychology, Neuroscience, and Reduction,* Wiley-Blackwell

Fallenstein, B.; Soares, N. (2014). Problems of self-reference in self-improving space-time embedded intelligence (Presented at Artificial General Intelligence: 7th International Conference, Quebec City)

- (2015a). Reflective Variants of Solomonoff Induction and AIXI (Presented at Artificial General Intelligence: 8th International Conference, Berlin)

- (2015b). Two Attempts to Formalize Counterpossible Reasoning in Deterministic Settings. Counterfactual Reasoning. Counterpossibles (Presented at Artificial General Intelligence: 8th International Conference, Berlin)

Feyerabend, P. (1962). Explanation, Reduction and Empiricism, in H. Feigl and G. Maxwell (ed), *Scientific Explanation, Space, and Time*, (Minnesota Studies in the Philosophy of Science, Volume III), University of Minneapolis Press, Minneapolis

Fodor, J. A. (1974). Special Sciences (or: The Disunity of Science as a Working Hypothesis. *Synthese* 28 (2):97-115

Goertzel, B. (2015). Superintelligence: Fears, Promises and Potentials. *Journal of Evolution and Technology* , Vol. 24 Issue 2: 55-87

- (2016). Infusing Advanced AGIs with Human - Like Value Systems : Two Theses. *Journal of Evolution and Technology*, Vol. 26 Issue 1: 50 - 72

Hull, D. (1972). Reductionism in genetics – biology or philosophy? *Philosophy of Science* 39:491–499.

- (1974). *Philosophy of biological science*, Prentice-Hall, New Jersey

- (1976). Informal aspects of theory reduction, in R.S. Cohen and A. Michalos (eds), *Proceedings of the 1974 meeting of the Philosophy of Science Association,* D. Reidel, Dordrecht

Kitcher, P. (1984). 1953 and all that: a tale of two sciences. *Philosophical Review* 93:335–373

Legg, S.; Hutter, M. (2006). A Formal Measure of Machine Intelligence. *Proc. 15th Annual Machine Learning Conference of {B}elgium and The Netherlands*, 73-80

- (2007). Universal Intelligence: A Definition of machine intelligence. *Minds and Machines,* 17 (4):391-444

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.; Veness, J.; Bellemare, M.; Graves, A.; Ridemiller, A.; Fidjeland, A.; Ostrovski, G.; Petersen, S.; Beattie, C.; Sadik, A.; Antonoglou, I.; King, H.; Kumaran, D.; Wierstra, D.; Legg, S.; Hassabis, D. (2015). Human-level Control Through Deep Reinforcement Learning. *Nature* 518: 529-533

Nagel, E. (1961). *The Structure of Science: Problems in the Logic of Scientific Explanation,* Brace & World, Harcourt

Putnam, H. (1963). Brains and Behavior. *Mind, Language, and Reality: Philosophical Papers*, Vol. 2: 325-341.

Rey, G. (1986). What's Really Going on in Searle's ”Chinese Room“, *Philosophical Studies*, 50: 169–85.

Russell, J. (2016). Google’s DeepMind chalks up AI landmark after beating Go world champion Lee Sedol. Retrieved April 29, 2016 from http://techcrunch.com/2016/03/09/googles-deepmind-chalks-up-ai-landmark-after-beating-go-world-champion-lee-sedol/

Searle, J. (1980). Minds, Brains and Programs, *Behavioral and Brain Sciences*, 3: 417–57

Soares, N. (2014). The Value Learning Problem. Retrieved February 2, 2016 from https://intelligence.org/files/ValueLearningProblem.pdf

- (2015). Formalizing Two Problems of Realistic World-Models: Solomonoff ' s Induction Problem. The Naturalized Induction Problem. Retrieved February 2, 2016 from https://intelligence.org/files/RealisticWorldModels.pdf

Strevens, M. (2016). Special Science Autonomy and the Division of Labor, in M. Couch, J. Pfeifer (eds) *The Philosophy of Philip Kitcher,* Oxford University Press, New York

Weinbaum, D.; Veitas, V. (2015). Open Ended Intelligence: The individuation of Intelligent Agents, *arXiv:1505.06366v2 [cs.AI]*

Winikoff, M.; Padgham, L.; Harland, J.; Thangarajah, J. (2002). Declarative and Procedural Goals in Intelligent Agent Systems (Paper presented at International Conference on Principles of Knowledge Representation and Reasoning, Toulouse)

Yudkowsky, E. (2001). Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures. *The Singularity Institute, San Francisco*. Retrieved February 2, 2016 from https://intelligence.org/files/CFAI.pdf

- (2004). Coherent Extrapolated Volition. *The Singularity Institute, San Francisco.* Retrieved February 2, 2016 from https://intelligence.org/files/CEV.pdf

Non-exclusive licence to reproduce thesis and make thesis public

I, Rao Pärnpuu,
(*author's name*)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

Ontology Identification Problem In Computational Agents,
(title of thesis)

supervised by Daniel Cohnitz,
(supervisor's name)

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, **02.05.2016**