Alma Mater Studiorum – Università di Bologna
in collaborazione con LAST-JD consortium:
Università degli Studi di Torino
Universitat Autonòma de Barcelona
Mykolas Romeris University
Tilburg University
e in cotutela con
University of Luxembourg

DOTTORATO DI RICERCA IN

Erasmus Mundus Joint International Doctoral Degree in Law, Science and Technology

Ciclo XXXII – A.A. 2016/2017

Settore Concorsuale di afferenza: 12H3

Settore Scientifico Disciplinare: IUS20

TITOLO TESI:

Foundations of an Ethical Framework for AI Entities: the Ethics of Systems

Presentata da: Andrej Dameski

Coordinatore Dottorato
Prof.ssa Monica Palmirani

Supervisori
Prof. Giovanni Sartor
Prof. Leon van der Torre

Esame finale anno 2020

Alma Mater Studiorum – Università di Bologna
in partnership with LAST-JD consortium:
Università degli Studi di Torino
Universitat Autonòma de Barcelona
Mykolas Romeris University
Tilburg University
and in cotutorship with
the University of Luxembourg

PHD PROGRAMME IN

Erasmus Mundus Joint International Doctoral Degree in Law, Science and Technology

Cycle XXXII – A.A. 2016/2017

Settore Concorsuale di afferenza: 12H3

Settore Scientifico Disciplinare: IUS20

THESIS TITLE:

Foundations of an Ethical Framework for AI Entities: the Ethics of Systems

Submitted by: **Andrej Dameski**

The PhD Program Coordinator	Supervisors
Prof. Monica Palmirani	**Prof. Giovanni Sartor**
	Prof. Leon van der Torre

Year 2020

uni.lu
UNIVERSITÉ DU LUXEMBOURG

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

PhD-FSTM-2020-68
The Faculty of Sciences, Technology and Communication

Alma Mater Studiorum – Università di Bologna, within the LAST-JD consortium:
Università degli studi di Torino
Universitat Autonòma de Barcelona
Mykolas Romeris University
Tilburg University

DISSERTATION

Presented on 06/11/2020 in Bologna
to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG
EN INFORMATIQUE

AND

DOTTORE DI RICERCA
IN LAW, SCIENCE AND TECHNOLOGY

by

Andrej DAMESKI
Born on 2 February 1987 in Skopje (Macedonia)

FOUNDATIONS OF AN ETHICAL FRAMEWORK FOR AI ENTITIES: THE ETHICS OF SYSTEMS

Dissertation defence committee

Dr Giovanni Sartor, dissertation supervisor
Professor, Università di Bologna & EUI (Italy)

Dr Leon van der Torre
Professor, Université du Luxembourg (Luxembourg)

Dr Emiliano Lorini
Senior researcher / co-head, Université Paul Sabatier (France)

Dr Juliano Maranhão
Associate Professor, Instituto de Estudos Avançados da Universidade de São Paulo (Brazil)

Dr Michał Araszkiewicz
Adjunct, Jagiellonian University (Poland)

To my beautiful, lovely girls—and you.

На моите прекрасни, мили девојки—и тебе.

—A.

Acknowledgments

Oh my, where do I start?

These past three and some years have been wild, exploratory, deep, and transformative beyond imagination. I've visited more places than I can count; I married my beautiful wife; got my two beautiful, beautiful girls; made so many, many new friends and acquaintances; strengthened the bonds with my close friends I already had; I even got to God/the Universe/the Dao/the Inexpressible/Nothingness/whatever-you-prefer; have seen people, systems, and pandemics come and go; and just simply had the experience and life I would not trade for anything.

On my path there were numerous people that deserve my deepest acknowledgments, respect and thankfulness. And I will start with none other than my own three families. First, a word of deep gratitude for my parents, Nikola and Elena, and my brother Damjan. They've made such a big sacrifice that only a loving parent or brother can make. They have supported me on the path from application, to rejection, to repeated application, to repeated rejection, to a complaint, and to acceptance. And when I finally got accepted, but was not given a scholarship, they have not hesitated at all to provide me all the financial and other support that I would need to undertake this project. I mean, sure—I worked an odd job or three in the meantime to be able to provide some money on my own. But the bulk of my scholarship money fell on their backs. For this I will be forever grateful and hope to repay it ten times over one day—both when they and my own children might be in need. Along the way we've had numerous debates, and we didn't always see an eye to an eye on many deep philosophical, political, or other questions; but we're a family, and that stands for something.

My 'second' family is the one I created myself by marrying my beautiful wife and having my (for now) two adorable girls. There's no right without obligation, and the gift of life I've received from my parents is an obligation as much as it is a right—to share the gift by creating new life that goes on after we're no longer here anymore. They have also sacrificed so much for this—days and weeks without me present, talking over video chat software, seeing the sadness in their eyes, and wishing I could hug them tight. But, as much as I went through all this for me, I also did it for them. My mission in life has always been to create a better world—and who do we create a better world than for our loved ones, our children, and for Being itself? Some years ago, while visiting the city-pearl of Ohrid in my beautiful Macedonia, I came across a deep and profound writing on a wall. It said: "We didn't inherit the world from our fathers; we borrowed it from our children". Now that's something we all have to ponder on really well. So, there are no words to express my thankfulness and respect for the sacrifice you've gone through for this, Tina, Mina and Dena (and another member of our family that maybe we are yet to meet). I hope I will be able to repay it, and that you'll always be by my side, as you are and always have been in my heart and mind.

My 'third' family is also one I created myself by loving, helping, appreciating, and being loyal to my group of close friends that I can't but consider my own brothers. The numerous debates we've had, the meditations we've done together, the long nights playing guitars, singing, listening to wonderful music, appreciating art, going to cinema or the theater, discussing, walking, working together, thinking, and just plainly sitting together doing nothing, enjoying the silent company. You and your brotherhood had such an effect on me and this work that I can't even begin to describe—but, again, hope to repay this debt tenfold. Aleksandar, Ognen, Filip, Petar, Martin, Ivan, Stefan, Bane—you are my beautiful, loyal people which have become a part of me that can't be removed without the patient dying.

I would like to sincerely thank professors Giovanni Sartor and Leon van der Torre, which accepted the role of my supervisors. Without you this wouldn't be possible. For most of the time they left me to my own devices, providing constructive (and, luckily, constructively destructive) feedback on what I've submitted, and have been very supportive (and busy) at all times. Thank you again. I thank professor Monica Palmirani, the director of this program, for her tireless dedication to get us through this all; as well as, Dr. Dina Ferrari, who has jumped numerous times to assistance in these shenanigans. I thank professors Massimo Durante and Ugo Pagallo for their hospitality and assistance during my great stay at the Università degli Studi di Torino. I also thank professor Antoni Roig Batalla for the same at the Universitat Autònoma de Barcelona. My great, great thanks for Dr. Livio Robaldo, Dr. Réka Markovich and prof. Leon van der Torre for their acceptance and assistance regarding my cotutele with the University of Luxembourg. I will always appreciate what you've done for me (especially since you didn't have to), and hope to repay it one day.

I would also like to thank professor Jordan B. Peterson, the Canadian clinical psychologist-turned-mentor of millions of people around the world, including me. Your words, especially the series on the psychological significance of the Biblical stories, on free speech and on archetypes, have resonated so deeply with me to the point of having a deep, transformative effect on my Being—and they even reflect in this thesis. Even though we are yet to meet in person, it's just like we already did. I hope to have the opportunity to sit down and have a chat with you sometime in the future, and wish you and your family all the best.

I thank my colleagues—Marco, Daniel, Rogers, Urbano—for your companionship along the way, as well as for your assistance and camaraderie. I also extend my thanks to all the people I've met on the way during these three years. Adriano, Antonio, Franz, Mlađo, Matteo, Borče, Nina, Bojana and Ognena, Sergio, Miloš, and many, many others—you are too numerous to all be included here, but I will always be grateful for your companionship.

I thank each and every place I have visited. The tradition and spirit of Bologna, beautiful grandiosity and endless sunny streets of Barcelona, Mediterranean flavor of Santander, Nice, Firenze and Torino, futuristic-old tradition of Dutch cities of Groningen, Rotterdam and Amsterdam, the freshness of Düsseldorf, the mysterious spirit of Luxembourg and Esch, the coffee of Italy, warmness of Spain (and Catalonia), and many, many more. I still walk my memories by browsing maps of all the places I've visited and all the people I've met.

I would also like to thank my own people and our country—Macedonians and Macedonia. We have taken (and still do take) a lot of abuse throughout the centuries, but we still managed to get on through and establish our own civilized and modern society while keeping our traditions. Many people are solely thankful to themselves for what they've achieved through life, but this is a very narrow and arrogant view. I would have never been able to achieve what I have if I wasn't growing alongside the beautiful and good people I was, and if I didn't get the education and everything else built and provided by the Macedonian people with blood and sweat. We had our ills, insecurities and immaturities, but we will push through—if we focus on being brave and good.

Before wrapping it up, I would like to acknowledge the *all* there is. All there is made this possible. It is both everything and nothing. We are all in this together, and we are all one, while being separate and unique simultaneously. The role of heroes is to serve life—its conservation, propagation and betterment. And we have the duty to make this all better—to create a better world. Finally, thank you.

Abstract

The field of AI ethics during the current and previous decade is receiving an increasing amount of attention from all involved stakeholders: the public, science, philosophy, religious organizations, enterprises, governments, and various organizations. However, this field currently lacks consensus on scope, ethico-philosophical foundations, or common methodology. This thesis aims to contribute towards filling this gap by providing an answer to the two main research questions: first, **what theory can explain moral scenarios in which AI entities are participants?**; and second, **what theory can explain the process of moral reasoning, decision and action, for AI entities in virtual, simulated and real-life moral scenarios?** This thesis answers these two research questions with its two main contributions to the field of AI ethics, a substantial (ethico-philosophical) and a methodological contribution. The substantial contribution is a coherent and novel theory named **Ethics of Systems Framework**, as well as a possible inception of a new field of study: ethics of systems. The methodological contribution is the creation of its main methodological tool, the **Ethics of Systems Interface**. The second part of the research effort was focused on testing and demonstrating the capacities of the Ethics of Systems Framework and Interface in modeling and managing moral scenarios in which AI and other entities participate. Further work can focus on building on top of the foundations of the Framework provided here, increasing the scope of moral theories and simulated scenarios, improving the level of detail and parameters to reflect real-life situations, and field-testing the Framework on actual AI systems.

Keywords AI ethics, ethics, philosophy of information, ethics of information, systems theory, general systems theory, systems science, ethics of systems, AI & law, agent-based simulations, digitization

Table of Contents

Chapter I. Introduction

1 Ethics of AI: On the brink of a new era in ethics

The Fourth Industrial Revolution is upon us. Everywhere we turn there is an electronic device that is either used or abused. We seem to be inseparable from our mobile devices, modern work is virtually unimaginable without some kind of a digital system, and people are using their brains at the highest intensity in history (while neglecting their bodies, often with terrible consequences). Everything seems to be turning digital and 'smart'—from smart houses, smart cities, smart contracts, digital assistants, communication, workplaces, relationships, and transportation; to news and propaganda, wars, killer drones, 'deep fakes', and personal data. The boundaries between the virtual and the 'real', between the digital and the analog are blurring with exponential speed—and most people struggle to even follow this breakneck pace, yet alone partake in it. Instead, they are left alone feeling isolated, with decreased agency, privacy and autonomy, and afraid for their livelihood further down in the future.

In the midst of all that we are also galloping towards introducing digital systems that can perform increasingly complex tasks in an automated and independent manner that is less and less supervised. What we now understand under the umbrella term of 'artificial intelligence' (hereinafter: AI) is a collage of technologies and tools that humans implement to perform tasks that are (presumably) boring, ineffective, obsolete, or too expensive to be performed by themselves. Sometimes, they are developed and implemented just for the sake of it, for the intellectual challenge; but possibly heading to devastating and unpredictable consequences. We are working hard on replacing humans with their more determined, cheaper, more efficient, and often less wise automaton counterparts. What can go wrong?

As with any development, there is the positive and the negative. Arguably, what people have been trying to do since the inception of our species—with various degrees of success—is to manage or avoid the negative while promoting the positive. There are 'various degrees of success' because sometimes we are not being able to devise good solutions to known problems in certain situations; and sometimes we do not even have a clear idea what the actual problem is. The latter may be even more dangerous than the former.

In the field of AI, the inability to agree on what 'artificial intelligence' actually means with a fair degree of precision might be the best indicator that we do not know what all the actual and potential problems are (going to be) arising from the widespread introduction of AI and automation in our societies. It is without a doubt that people in academia, government, and all around different spheres of human existence are working hard on this problem.

The European Union already undertakes significant steps in the direction of AI ethics (i.e. by forming the AI High Level Expert Group and the European AI Alliance; and there are also other independent projects such as CLAIRE). Across the ocean in the United States the debate is also lively, with the State Department, the White House, other governmental bodies, and private entities (e.g. Google, Microsoft, OpenAI) also joining in the debate. At the United Nations, the director of the UNESCO recently published an official article in the UN Chronicle (Azoulay, 2018). And of course, the Big Four (PwC, Delotte, KPMG, and EY) are all rushing forward into employing AI in their work, but at the same time publishing opinion pieces on the impact the technology will have for the years to come (Fagella, 2019). We have also for a while now been creating motion pictures (i.e. *The Matrix*, *Deus Ex Machina*, *The Terminator*, *2001: The Space Odyssey*) and science fiction (i.e. *Dune*, *With Folded Hands*) where the doom of reckless introduction of machines and algorithms into our world has been described with terrible vividness. To be fair, sometimes, but rather rarely, our AI overlords save us from ourselves (such as in Stanisław Lem's *Golem XIV*).

While fiction can be accused as being just that—fiction—it is a valid vehicle through which we explore our fears of new technologies and ways of existence, and try to devise methods to manage them positively. And with AI, there are plethora of potential and actual issues that we need to figure out how to manage before it gets deployed in a such a widespread manner that this becomes virtually impossible.

Some of these questions are about responsibility and accountability of automated systems, the loss of autonomy and agency of people, and their increase for automated systems, the loss of privacy and the abuse of personal data, the delegation of increasingly important tasks and roles to automated systems, the increasing complexity and opaqueness of these systems, their effect on the respect of human rights and dignities, whether these systems can be trusted even when we do not and cannot know what exactly they are doing, the perpetuation and exacerbation of bias, whether we as humans should remain the focus of moral theories, and many, many more. It is obvious that the list of issues on the table now or in the near future is huge, difficult, and complex.

In this work I am making a contribution towards identifying and managing the issues that (will) arise from the aforementioned developments, particularly in the domain of ethics, and particularly in the field of applied AI ethics i.e. moral scenarios in which AI entities participate. Although there is an increasing focus on the problems and details of the introduction of AI and automation in our societies, the efforts at solving them are various and diverse, and a bit 'lost'. The field of AI ethics currently lacks consensus on scope, ethico-philosophical foundations, or common methodology.

In a sense this is to be expected. As with any new development, we are struggling to figure out (if we can at all) what the exact issue is, how to model it, what its attributes and parameters are, where it might take us, where we *want* it to take us, and how to get there. In this manner, I believe that some of this work in the ethics of AI domain should be *foundational*—in the sense that it will enable us an underlying framework of thinking and acting in a morally-sound and sustainable way long into the future.

This is the exact focal point of this text. My thesis is focused on two research endeavors. The first is offering the possible foundations of a comprehensive ethical framework for AI entities. The purpose of this Framework is to provide the means to formally explore m**oral scenarios in which AI** entities are included as participants. The second is testing and demonstrating the capacities of this framework by exploring hypothetical moral scenarios.

The name I have given to this framework is the **Ethics of Systems Framework** (hereinafter also: the Framework[1]). Additionally, the Framework itself should be foundational, coherent, contextual, computationally- and logically-representable (in principle), ethically and morally sound, and be implementable in real life scenarios in which AI systems and humans participate. It should offer a systematic approach towards better moral outcomes. It should accommodate modeling of moral scenarios with high precision and quality, adaptation to changes in context, and deliver suitable and morally-sound solutions.

Needless to say that this is not an easy task at all. I am not purporting to have figured out the 'Holy Grail' of AI ethics. What I am merely trying to convey is an insight into ethics in general, and particularly in ethics of AI, that I believe can significantly contribute towards the whole effort. However, since forming and testing a comprehensive theory is the work of decades (centuries, even?) and way beyond the scope of a doctorate, I have settled on formalizing the *foundations* of such a framework.

[1] Any other frameworks will be duly defined and used in different textual configurations.

I try to address as many relevant issues as I possibly could, and then distill the solution (the new theory) in as simple and as foundational manner possible. My effort aims to contribute positively towards the prediction, understanding, and management of the problematic morally-burdened effects coming forth. And with that, I hope to contribute towards my main motivation: bringing upon a (morally) better world.

2 Purpose of the study

The purpose of the study performed here is to devise the foundations for a comprehensive (meta)ethical framework applicable to moral scenarios in which AI entities are participants; and to test the framework in hypothetical scenarios. It is split between three sub-purposes: 1. a general contribution to AI ethics and law; 2. exploring implications for AI ethics from systems theory and ethics of information; 3. and as possible cybernetic implementation into AI entities.

2.1 General contribution to the field of AI ethics and law

The first sub-purpose of the study is to contribute to the field of AI ethics in general. This is performed as an attempt at reconciling the dominant ‘classic’ ethical theories of today (i.e. deontology, teleology, virtue ethics) and some newcomers (i.e. ethics of information, environmental ethics and deep ecology, feminine ethics, ethics of care) into a singular, unified framework. The Framework thus renders the aforementioned into special cases of itself, depending on context and aims.

The intent is that this reconciliation will provide a comprehensive *system* of ethical principles, methods, and practices of reasoning. This system should be flexible, adaptable, iterative, simple yet expandable by and for all potential moral reasoners (both human and AI entities), and applicable for a wide variety of moral scenarios.

2.2 Implications for AI ethics from systems theory and ethics of information

However, in order to make a strong attempt at such reconciliation, there is a need to discover the foundations of ethics applicable to the participation of AI entities in human societies. These foundations can be discovered in areas of study whose implications give rise to the foundations of ethics and ethics of AI.

With this in mind, a decision was made to dive into scientific disciplines behind (hence “meta-”) ethics: into systems theory and ethics of information (which itself emerges out of philosophy of information). In this sense, another purpose of this investigation is to provide a study of the implications derived from these meta-ethical disciplines, how they give the rise to the (meta)ethical foundations of the Framework, and how they contribute to ethics of AI in general.

2.3 Potential cybernetic implementation of the Framework in AI entities

The third sub-purpose of the study is the potential cybernetic implementation of the Framework in AI entities. The Framework aims to enable better understanding, modeling, and solving moral scenarios in a morally-sound manner, by and for AI entities.

The groundwork mentioned above is based on axiomatic ethical foundations which are computationally representable in principle, and can be implemented into AI systems through engineering and programming means. With the help of its axiomatic foundations the whole ethical Framework is designed as a system, which can then be implemented, partially or fully, as a cybernetic subsystem in an artificial entity. These capacities are tested and demonstrated in the second part of the thesis.

3 Research objectives and questions

3.1 Research question(s)

3.1.1 Main research questions

As is mentioned in the introduction, there is a need to develop a unified approach when dealing with ethics of AI. This approach should be focused on understanding and modeling moral scenarios where AI entities are involved. But it also should be focused on enabling and aiding the moral reasoning of AI entities in these scenarios, so that we can improve their behavior in a morally-relevant manner.

We need to understand what are the elements (and their attributes) that compose moral scenarios which include AI entities; such as the participants themselves, moral processes, moral reasoning and (non-)acting processes, conflict, distribution of resources, positive and negative morally-burdened effects, and similar issues.

Unfortunately, no such adequate comprehensive theoretical, methodological or practical approach exists which both contributes on an ethical (substantial) and formal level in regards of modeling and managing moral scenarios in which include AI entities participate (see 5 Ethics of AI in Chapter II. Literature review). Hagendorff (2019) performed a wide review of ethical guidelines for AI ethics, and identified that these kinds of guidelines are rarely implemented or followed—either by the engineers and designers of AI systems, or by the AI systems themselves. Rarely such provide formalized methods to instantiate moral reasoning based on the ethical principles they offer. One such attempt is Floridi's Ethics of Information (Floridi, 2013), but it is still rudimentary and is based on the informational level of abstraction, hence is not yet adequate and mature enough. Systems theory and its own systemic level of abstraction, on the other hand, can provide a significantly improved explication of AI moral scenarios.

With this in mind, the main research objective is to explore the feasibility and utility of developing and testing a theory (in the form of a (meta)ethical framework), based and building upon on both systems theory and ethics of information, that can successfully explain, model, and help solve moral scenarios in a morally-sound manner, by and for AI entities.

Therefore, the following two research questions are addressed:

- **What theory can explain moral scenarios in which AI entities are participants?**
- **What theory can explain the process of moral reasoning, decision and action for AI entities in virtual, simulated and real-life moral scenarios?**

A candidate theory that provides a satisfactory answer to these questions would be able to formally represent moral scenarios (and their components and attributes) in which AI entities participate, and provide the means to perform moral calculus on available courses of (in)action in accordance with formally-designed moral theories.

3.1.2 Research sub-questions

For the purpose of answering the above two research questions, there is a number of sub-questions that need to be answered. Each one of these pertains to key aspects of the theory laid in this text, and naturally emerge out of the main research questions above.

- ○ What are the major ethical issues raised by the introduction of AI entities in the world and in human societies?

- What are the foundational systemic and informational attributes of moral scenarios whose participants include AI entities?
- Which are the foundational ethical principles, concepts and methods of reasoning relevant to AI entities?
- Can the foundational ethical principles, concepts and methods of reasoning relevant to AI entities be systematized into a coherent (and possibly comprehensive) ethical framework?
- In what way can such ethical framework provide means for, or assist, reasoning in moral scenarios in a morally-sound manner?
- Can such an ethical framework be translated or paraphrased into legal, technical, engineering, and other instruments?
- What are the ethical, scientific, and possibly legal implications that this kind of a comprehensive study brings on AI ethics, and ethics in general?

3.2 Expected outcome of the study

The expected outcome of this work is the discovery of the foundational ethical principles of moral scenarios where AI entities are participants. The further outcome is using these foundational principles to design the skeleton of a comprehensive ethical framework for AI entities as a modeling, reasoning and (in)action-informing system. And the final expected outcome is to successfully test and demonstrate these capacities of the framework in hypothetical scenarios.

My intent is to make a compelling case that:

1. Ethics of AI is a subject of study that is based on a unifying set of foundational ethical principles;
2. These ethical principles can be discovered and put before scientific inquiry;
3. These ethical principles can inform the design of ethical theory that can enable study of the participation of AI entities in moral scenarios;
4. The design of such ethical theory can inform the practical design of actual machine (sub)systems that will govern the participation of the AI entities in moral scenarios in a morally-sound manner; and,
5. This text is a scientific inquiry that has resulted in such findings (as those mentioned above), and offers a compelling ethical theory and framework for AI entities.

4 Scope

The scope of the study is limited on analyzing ethical perspectives primarily in the domain of AI ethics and *AI entities* (agents and patients). The scope is thus limited in this manner even though the Framework is applicable on the whole plethora of systems in general, as we will later see in Chapter III. Towards Ethics of Systems (the Metaethics).

However, there is also the necessary contextual incursion into non-AI entities (such as humans, the environment, animals, ecosystems, and similar). The reason is simple: ethics of AI is essentially about ethics in general, which also encompasses human-centric ethical research. When discussing AI ethics we are necessarily discussing participation of AI entities in moral scenarios that typically include humans and/or some of the systems we have either created or participate in—institutions, ecosystems, states, organizations, simulations, and others.

Additionally, AI ethics is (for now) a human endeavor, and we are interested in managing the morally-burdened effects of the widespread introduction of AI and automation into our societies. A case may be made that we even hold a certain human-centric bias when working on these issues, a bias that has at least one particularly compelling reason to exist: we would like to avoid serious adverse effects to humanity, human civilization, human societies and individual humans that may be caused by the introduction of new and immensely powerful technologies. These adverse effects typically happen when we avoid the responsibility to explore all the contentious moral, ethical and practical issues arising with their utilization.

This limitation thus imposed on the scope of study is done for the purpose of making the research feasible and narrow enough so that it fits the typical format of a doctoral thesis work. However, by analogy and through contextual interpretation the findings, the study—as well as the framework presented—can be translated into and hence applied to other ethical contexts.

5 Research methodology

The study I present here is multidisciplinary and predominantly qualitative. It draws upon insight from systems theory, theory of information, ethics, law, artificial intelligence, and other fields. Needless to say, this results in the utilization of a multidisciplinary methodology that comes about from the different discussed domains.

However, this diverse methodology is used in a manner that is consistent with the text itself. This at times necessarily results in using only some particular elements of a certain field's methodological set that are relevant. It also results in an approach with varying levels of depth, at times ignoring deeper methodological tools or expositions that were not relevant for the discussion itself.

The study uses 3 main methodological approaches: (general) systems theory, grounded theory, and information ethics. These are aided by other methodological approaches that fit particular components.

5.1 Systems theory methodology

The main methodological toolset that is used throughout this text is the one of (general) systems theory. This approach is chosen because it is versatile, adaptable, holistic, and harmonizing—which is exactly the aim of the study itself. There are several key concepts discussed in (general) systems theory: emergence, hierarchy, communication and control (cybernetics). These, and other important ones (such as basic set theory, entropy, structure) are used as the lens through which I will conduct my research. The aim is, of course, to discover the various systems and structures that are relevant for AI ethics, and, at the same time, *systematize* the findings into a coherent whole (the ethical framework itself).

To quote (Klir, 2001), "Since at least the beginning of the 20th century, however, it has increasingly been recognized that studying the ways in which things can be, or can become, organized is equally meaningful and may, under some circumstances, be even more significant than studying the things themselves [...] While the systems perspective was not essential when science dealt with simple systems, its significance increases with the growing complexity of systems of our current interest and challenge". Furthermore, Kanungo and Jain (2007) also state that: "From a methodological standpoint, systems theory (the field) and systems thinking (the worldview and approach shared by those who subscribe to systems theory) can help IS *[information systems]* researchers frame and address complex and messy problems".

(General) systems theory methodology is a particularly suited methodological approach for the type of study conducted here, and it ties well with the second main methodological tool—grounded theory.

5.2 Grounded theory

The second main methodological tool that is used in this study is grounded theory. Grounded theory is an inverse methodological approach, where the researcher first gathers data and then attempts to devise an abstract theory that explains a certain phenomenon, process, or relation (or a set of these). To quote from Creswell (2014), "Grounded theory is a design of inquiry from sociology in which the researcher derives a general, abstract theory of a process, action, or interaction grounded in the views of participants. This process involves using multiple stages of data collection and the refinement and interrelationship of categories of information (Charmaz, 2006; Corbin & Strauss, 2007)".

Having into mind that the end goal of this study is the development of an ethical framework for AI entities, the general style of grounded theory fits the whole approach. As mentioned before, it also ties neatly with the systems theory approach outlined above, as the very construction of a theory is an *emergent* result.

Of course, the whole process of discovery, data gathering and research is a constant travel between data gathering, analysis and synthesis. The data gathering in this case would be performed as a systematic and informed extraction of the data out of texts and discussions in the fields included in the study e.g. ethics, law, AI science, philosophy, philosophy of information, systems theory. Information will also be gathered from theoretical and actual studies performed with AI entities (e.g. MIT's Moral Machine). These will then be consolidated into a theory that attempts to explain the basic 'social' process of AI ethics (Salkind, 2010; p. 552).

5.3 Information Ethics methodology

The third most important methodological tool that is used is (some parts of) the information ethics methodology, namely the method of level(s) of abstraction and the object-oriented model of moral action.

5.3.1 Method of Level(s) of Abstraction

This method is used when analyzing the various systems that appear in AI ethical scenarios, so as to provide a method of discerning the different components of the scenarios. Systems can be components of other systems, and be comprised of multiple systems themselves. The discernment of these configurations ("our view of the world", to quote Floridi) can be helped by using the method of abstraction levels (Floridi, 2013).

Another contribution this method makes is the choice of the systems theory approach outlined above, as it was determined that the systemic level of abstraction provides the best unifying research viewpoint with which to explore the AI ethics problematic.

5.3.2 Object-oriented model of moral action

The object-oriented model of moral action (Floridi, 2013) is a methodological tool that is used in this text in multiple manners. Firstly, it is used to conceptualize moral scenarios, which can then be analyzed successfully using the other methodological tools we have at our disposal. Secondly, it is used to provide the way to systematize the whole ethical scenario and its components i.e. to show how it is a *system in itself*.

5.4 Moral reasoning and communication

Of course, if we are to conduct a well-formed research on the behavior of AI entities in moral scenarios, we have to pay attention to the moral reasoning and communication that take place in them. Moral reasoning is part of a the general study of *reasoning* as a cognitive capacity. For this purpose, some basic elements from argumentation theory and decision theory (see below) will be used, to be able to analyze how moral reasoning happens, and how it *ought* happen given specific goals and contexts in mind.

5.4.1 Argumentation theory

As mentioned above, argumentation theory is consulted as an auxiliary approach when studying moral reasoning and communication that takes place in moral scenarios. Although at times it might seem as too rigid a framework to be used to analyze the dynamic and fine-grained world of moral discourse, it provides a great tool-set to conceptualize the actual cognitive moral process of the entities in moral scenarios, the (moral) arguments they provide to themselves and others for particular courses of (in)action, and the way these can be modified, defeated or overridden—thus changing the course of (in)action and the path in which the moral scenario develops.

5.4.2 Decision theory

Similarly to argumentation theory, decision theory is consulted as an auxiliary approach in an identical context. Authors in this field have developed methodological tools and approaches that can prove useful for the work that I am performing here.

5.5 Literature review

And finally, an extensive literature review is performed in the subjects of AI ethics and ethics in general, law, AI law, systems science, philosophy and ethics of information, and other related fields. The literature under scrutiny will be primarily recent work (i.e. work published in the past 15 years). However, there will also be an occasional necessary ingress into older work that has significant impact on the presented findings.

6 Thesis contributions

We are currently far from reaching a consensus on the basic building blocks in the field of AI ethics, such as scope, ethico-philosophical foundations, and common methodology. The contribution of this thesis is exactly in this direction.

It is making a significant contribution on a substantial level with the Ethics of Systems Framework itself, with insight gathered from ethics, AI ethics, systems theory, and philosophy and ethics of information. This is aided by the capacity of the Ethics of Systems Framework and Interface to bridge the gap between form and substance. This thesis is also making a significant contribution on a methodological level by delivering the Ethics of Systems Interface. The Ethics of Systems Interface is a methodological tool that can be used to explicitly and formally represent moral scenarios in a consistent and coherent manner, paraphrasable and implementable across disciplines, authors, and organizations.

7 Structure of the thesis

The writing is progressive and builds upon previous findings throughout the text. The reason the text was structured in this manner is to provide story-like flow to the findings that logically follow from previous conclusions and emerge as insights to you, the reader.

For example, each chapter where a personal research contribution is made (Chapters 3 and 4) contains a section with implications that are carried forward to the next chapters (and for future work, as with Chapter 5).

The text is split into 7 chapters.

This, Chapter 1, is the introductory one.

Chapter 2 is focused on literature and state of the art review. This is where I dive into available literature on the subject of AI ethics, primarily focusing on recent work (as already mentioned above i.e. work published in the past 15 years); with the occasional necessary ingress into older work that has significant impact on the presented findings. The rationale is that ethics is an old subject, old as civilization itself, and much research done (and many of the dominant theories) in the field dates way back before the current times—but cannot be considered as obsolete at all!

Chapter 3 contains the first half of the research effort. It is using a multidisciplinary approach (i.e. systems theory, theory of information) to discover the meta-ethical principles that apply to moral scenarios which include AI entities. Through this, the foundations of a comprehensive ethical framework for AI entities—the Ethics of Systems Framework—and of its main methodological tool—the Ethics of Systems Interface—are set. The foundations are a common set of axiomatic ethical principles that are computationally representable, and that can enable understanding, modeling, and solving moral scenarios in a morally-sound manner.

Chapter 4 is the second half of the research effort. It is focused on exploring two moral scenarios, the classic Trolley problem, and the Trust and Trade scenario (a turn-based trading simulation). This is done by designing four classes of ethical theories (consequentialism, deontology, virtue ethics, and EoS four ethical principles), in total 8 moral theories. The theories are then applied and tested in the two scenarios, whose development is then tracked. The purpose of this chapter is to demonstrate that Ethics of Systems Framework and Interface can handle various moral situations with which it might be tasked with managing.

The next Chapter 5 is a discussion on the substantial (ethico-philosophical), technical, and scientific implications that arise from the Framework. This is an explication of the consequences arising from the conclusions reached in the meta-ethics and ethics part of the thesis text. This is also the chapter where I discuss the potential future research work on the subject that arises from the presented findings.

And finally, Chapter 6 holds the conclusion. It is where I recapitulate what is the outcome of my research work, and conclude that I have made a compelling case for the design of the foundations of a comprehensive ethical framework for AI entities.

Chapter 7 is a container for the bibliography consulted and cited throughout this research.

Chapter II. Literature review

1 Introduction

This section is an opportunity to dive deep into the relevant literature regarding the subject matter. Since this is a predominately qualitative research effort, the literature review serves as a point of entry into the qualitative aspects of AI ethics. The end goal is to gather and analyze enough information on the subject matter which will be synthesized into a coherent result that can stand under the scrutiny of time and reviewers.

The literature review follows the structure of the thesis text, and is thematic. It starts with an overview of the important literature in metaethics. Then it follows to general ethics, before diving deeper into literature specifically focused on AI ethics; and will finish with an overview of relevant international documents.

Along this route, I will identify key concepts who will be included in Appendix II. Key concepts. These concepts represent the most important ethical considerations that the Framework will have to provide conceptualization for in moral scenarios.

2 Metaethics

2.1 Systems theory

Systems theory or systems science[2] is a promising approach for ethics in general and AI ethics, but one that has rarely been used in research work. The probable reason is that researchers within the humanities, ethics, law, IT and other related fields are not familiar with it enough, not even with its basic tenets.

Before we go into how systems theory approach can be applied to AI ethics, however, we ought to explore what systems theory is. Simply stated, systems theory is

> "… that field of scientific inquiry whose objects of study are systems" (Klir, 2001; p. 3).

However, this definition is immediately pointed out by Klir in the same text (Klir, 2001) as not sophisticated enough. The reason is that systems theory is centered on the property of *systemhood* as its main focal point, using the auxiliary properties of *thinghood* and *setness* to help in its inquiry. Thus, the definition that Klir suggests in the text is the following:

> "Systems science is a science whose domain of inquiry consists of those properties of systems and associated problems that emanate from the general notion of systemhood" (Klir, 2001; p. 5).

Systems science has traditionally been approached in the formal sense as a subject of mathematics and mathematical analysis, set theory, and logic. Some additional 'traditional' incursions of research have been performed in computer science (i.e. the analysis of distributed systems, or computer systems in general), architecture, and in military analysis (Skyttner, 2005).

However, since the notion of a *system* is not reserved only for the engineering- and mathematically-minded researchers, there have been some incursions of utilizing systems science in social sciences, administration, ecology (i.e. eco*systems*), economics, law, politics, and similar ('soft'?) fields of study; even though the notion that a body of knowledge of any 'science' ought to be *systematized* is old as science itself.

[2] The terms *systems theory* and *systems science* are conflated and used interchangeably for all practical purposes of this text.

Probably one of the best introductory volumes on the formal aspects of systems science is George J. Klir's *Facets of Systems Science* (Klir, 2001). It introduces what systems science is; what a system is; its common-sense and formal definitions; properties of systems; formal approaches to manipulation of symbols representing systems, their components and their properties; systems science methodology and metamethodology; complexity; and plenty more. Many of these concepts will be explored in Chapter III. Towards Ethics of Systems (the Metaethics), where I go deeper into how systems science is applicable to the subject matter of this thesis.

However, the second-best contribution of Klir's book is the inclusion of important and influential papers in systems science from all areas of study, not just formal or mathematical. These include renowned papers of authors such as von Glasersfeld (2001), Maturana, Varela and Uribe (Mingers, 1991; Varela, Maturana & Uribe, 2001)(Varela et al., 2001), Ashby (2001), Prigogine (2001), and of course, Zadeh (2001). The breadth of research is wide and varied, from analyzing general systems theory, social systems, complexity, cybernetics (control in systems), autopoiesis (self-creating systems), information and its laws, economics and systems theory, systems science methodology, and many more.

Of course, there are other researchers mentioned but not fully included in the book mentioned, which are highly influential in the field, such as Niklas Luhmann (2013), Ludwig von Bertalanffy (1969), Haken (1983), and Rosen (1978). Luhmann is probably the most well-known of these, by getting deep into social systems and autopoiesis, providing insightful analysis why and how social systems form, how they function, what is their purpose, and what precisely are systems. In this respect, in his book *Introduction to Systems Theory*, he offers a significant contribution to systems science by providing a *negative* definition of a system, in the sense of defining what a system *is not*. He states that a system is the *difference* between the system and its environment (Luhmann, 2013; p. 44) i.e. the system is *not* the environment around it (see more in Chapter III. Towards Ethics of Systems (the Metaethics)).

One thing missing from the whole systems science research endeavor is the debate between systems *constructivism* and systems *realism*. Although they are mentioned shortly in Klir's book, there is a lack of further attention on this issue (apparently as an attempt to avoid dabbling in tricky epistemological issues), and a glaring lack at mentioning literature that supports systems realism. I will visit this debate in Chapter III. Towards Ethics of Systems (the Metaethics) with the purpose of giving a brief spotlight to this issue that can make or break whole theories and methodological approaches. In that respect, there might be a third 'middle' way to approach the problem i.e. a so called *constructivist realism* (or similarly titled) (Cupchik, 2001).

In general, the field of systems science provides a new kind of view in study—a holistic one. Whereas 'traditional' science has been dealing with a more analytic mindset, systems science attempts to provide a unifying study of phenomena (systems) that explains the 'big picture'. This holistic approach can be applied to studying all and any kind of systems, including ecosystems, societies, groups, individuals, families, and their systems—financial, moral, legal and other systems. This is the main reason why I chose systems science approach for this research.

2.2 Philosophy of information

Philosophy of information is a research approach that has (relatively) recently developed into a full-blown and comprehensive theory on data, knowledge and information. Its domain is comprised of subjects such as what is information; the relationship between physics, 'physicality,' and information; computation; algorithms; semantics and syntax; abstraction and levels of abstraction; logic of information; cognition; value of information; information-theoretic philosophy of mind; integrated information; and in general, the study of all subjects stemming out of, or otherwise related to, information.

Notable authors in this field include Patrick Allo, João Antonio de Moraes, Fred Adams, Phyllis Illari, Federica Russo, Stephen Rainey, Mariarosaria Taddeo, David Gamez, Christoph Szhultz, Giuseppe Primiero, Laszlo Kosolosky, Rafael Capurro, Giulio Tononi, Christof Koch, Bert Baumgaertner, Luciano Floridi, and plenty others (Floridi, 2016a) (Tononi, Boly, Massimini & Koch, 2016) (Baumgaertner & Floridi, 2016).

One of the front-runners of this relatively young discipline is prof. Luciano Floridi with his incursions in philosophy of information (Floridi, 2011), ethics of information (Floridi, 2013), and recently, logic of information (Floridi, 2019). It seems fitting to use his definition of philosophy of information as our working definition in this work. Thus,

> "(D) philosophy of information (PI) $=_{\text{def.}}$ the philosophical field concerned with (a) the critical investigation of the conceptual nature and basic principles of information, including its dynamics, utilisation, and sciences, and (b) the elaboration and application of information-theoretic and computational methodologies to philosophical problems" (Floridi, 2002).

A question naturally arises in this context: **how is philosophy of information relevant to ethics of AI, and even more precisely, to this work?**

This question can be answered in a three-faceted response.

On one hand, in the section on systems science I discuss how systems (entities) act in the world, basing their deliberations and decisions on data and information. These include data and information *about other systems* and *about themselves*. This is especially relevant to ethics, since moral reasoning and behavior is in large part focused on entities deliberating and acting out in the world in regards of other entities, while pursuing their personal goals. Moral life is a "highly information-intensive activity" (Floridi, 2013; p. 20).

On the second hand, consciousness, cognitive capacity and awareness are very tightly related to morality, and thus to AI ethics. For example, what is defined as *dignity* in Floridi's work is dependent, at least in part, on the cognitive capacity of an entity. The argument is that stronger cognitive capacity can, in turn, provide stronger capacity to know (and be aware of) the infosphere, and take care of it (Floridi, 2013; p. 76). Typically, stronger cognitive capacity translates to 'stronger' consciousness, and sometimes – self-consciousness (consciousness of second order). This is reminiscent of C. S. Lewis' take who considers that

> "[T]he more a Being is made from a better material, the smarter, stronger and freer it is, that much better it would be if it goes in the right way, but also that much worse if it goes in the wrong way. A cow cannot be neither too good nor too bad, a dog can be better and worse; in a larger measure can a child be bad or worse; in a larger measure still can a typical grown up man be better and worse, and in a larger measure still a genius; a superhuman spirit can, in turn, be the either the best or the worst of all" [translation from Macedonian mine] (Lewis, 2017; p. 75; Macedonian translation).

Cognitive capacity and consciousness also tie in to the notion of *integrated information* introduced by Oizumi, Albantakis and Tononi (2014). We will see further down the line that this perspective enables making an attempt at developing moral calculus, which can prove very useful for AI entities. It also gives some moral implications (e.g. about the moral 'veil of ignorance') that may inform how AI entities ought to be designed and implemented, and how information served to them ought to be presented, to enable the most favorable moral choice in a practical moral scenario.

On the third hand[3], a whole sub-field of both ethics and philosophy of information has emerged, with the title of *ethics of information*. It strives to provide a fresh perspective on ethics, one based on the aspects of information. This also, and again, may include the notion of *integrated information* as one approach to

[3] I know, that's one too many hands, but probably not for a robot?

measuring the Being of an entity. The attempt is to provide a significantly better explication of moral scenarios and moral processes in which (informational) entities are participants; and to stimulate finding the best possible moral choice in the context.

Since ethics of information is the second major approach I am utilizing in this work, I am including a separate section (please see immediately below) that is focused on its literature review.

2.2.1 Ethics of information

Ethics of information spun out of philosophy of information, with the purpose of applying and further developing new insights gained into the field of ethics. It draws upon findings and concepts such as levels of abstraction, the infosphere (an informational-theoretic representation of the world), informational entity (an information-theoretic representation of entities in the world), entropy and deflationary effects it causes on informational entities, informational structural realism, information-theoretic notion of Being, and other; and further explores them in the field of ethics, thus offering a fresh perspective.

A working definition for this work would be thus, that ethics of information has a

> "... role as a macroethics, that is, as an ethics that concerns the whole realm of reality, at an informational level of abstraction" (Floridi, 2013; p. 27).

Ethics of information offers innovation in ethics on both the ontological and epistemological level. For example, a novel description of the world as the infosphere, where entities are represented as informational entities (based on the informational level of abstraction), while incorporating entropy as the ultimate proximal cause of moral bad and evil, and offering the concept of initial moral worth and dignity to all entities, is an approach that offers a strong contribution from an ontological perspective. Additionally, since it stems from philosophy of information it also has something significant to contribute in an epistemological sense, especially in the fields of perception and communication— contributions starting even from the mid 20th century, notably with Shannon's *A Mathematical Theory of Communication* (Shannon, 1948).

There is a plethora of prominent authors in ethics of information, some of which naturally 'spill over' from philosophy of information. These include (and are not limited to) Charles Ess (2009), Mariarosaria Taddeo (2017), Brent Mittelstadt, Allo, Taddeo, Wachter and Floridi (2016), Patrick Allo, Sandra Wachter, Ugo Pagallo (2017), Massimo Durante (2011), Rafael Capurro (2006), and Floridi (2013). Floridi has dedicated the second book in his magnum opus trilogy precisely on ethics of information (Floridi, 2013), developing a fresh theory and promoting the informational worldview in ethics. This worldview comprises the second most important perspective that I use in this work, and fits nicely with the dominant one—the systems worldview.

Capurro has criticized Floridi's ontological approach in general, and in the field of ethics. I will account for these contentions when diving deeper into ethics of information in Chapter III. Towards Ethics of Systems (the Metaethics).

3 Moral reasoning

There can be no serious ethical theory that does not propose at least the basic tenets of how moral reasoning ought to take place in moral scenarios. There are plenty of approaches purporting to explain moral reasoning in context. However, when dealing with AI such moral reasoning can be augmented by drawing on findings from already established reasoning theories, such as value theory, argumentation theory and decision theory. Thus I have included auxiliary readings in these two topics.

3.1 Value theory

Ethics and morality are focused on what is valuable and what is right, their opposites, and the fine-grained positions in-between. In general, when dealing with ethics we are trying to protect, conserve and improve what is of value. In consequentialist theories this is known as ‘the Good’, in deontological theories as ‘the Right’, in virtue ethics as ‘the Good Life’, and so forth. Each ethical theory has as its own object *something* (e.g. a particular state of matters, duty, (in)action, etc.) that is axiomatically valuable[4]; and methodology how to achieve, improve, or protect it.

Thus, whenever we are dealing in ethics and morality, we need to proclaim *something* as having an axiomatic, a priori (intrinsic) value. This *something* will serve as a compass to guide our moral reasoning i.e. to figure out what is (morally) good and right, what is bad and evil, and what is *more* or *less* of the aforementioned (see Table 1: Value modalities below for the different value modalities; note that the qualifiers *positive* and *negative* imply the word *value(able)*). Without having accepted *something* of a priori value our moral compass would be confused, pointing at everything and at nothing—thus leaving us morally lost. This is equally applicable when dealing in AI ethics.

Table 1: Value modalities

Moral value concept	**Morally positive**		**Morally neutral**	**Morally irrelevant**	**Morally negative**	
	superlative	*comparative*			*comparative*	*superlative*
The Good	the Good / the best	better	neutral / equally good and bad	amoral / irrelevant	moral bad / worse	evil / moral worst / moral catastrophe
The Right	the most right / imperative	the right	neutral	amoral / irrelevant	the wrong	the absolutely wrong
Virtuosity	most virtuous / virtue	more virtuous	neither virtuous nor vicious / average(ly) virtue/ous	amoral / irrelevant	more vicious	most vicious / vice
etc.						

3.1.1 The value spectrum

Our moral intuitions include a plethora of value categories. We can commonly recognize states, decisions and/or (in)actions that sit and move somewhere on a spectrum[5] from absolutely positively valuable to absolutely negatively valuable. And in the middle there is often a certain space recognized for the morally irrelevant or neutral. We also intuitively accept the role agency, intentionality, and capacity have over where the above states or decisions or (in)actions sit or move on the spectrum. And finally, we take the effect of time as a very important moral consideration. Illustration 1 shows this spectrum in a simplified form.

[4] Thus, *axiology* is study of goodness or value in the widest sense of the terms.

[5] Although some moral philosophers deny that moral phenomena can be represented in the manner I offer here. See section 3.1.3 below for more commentary on this issue.

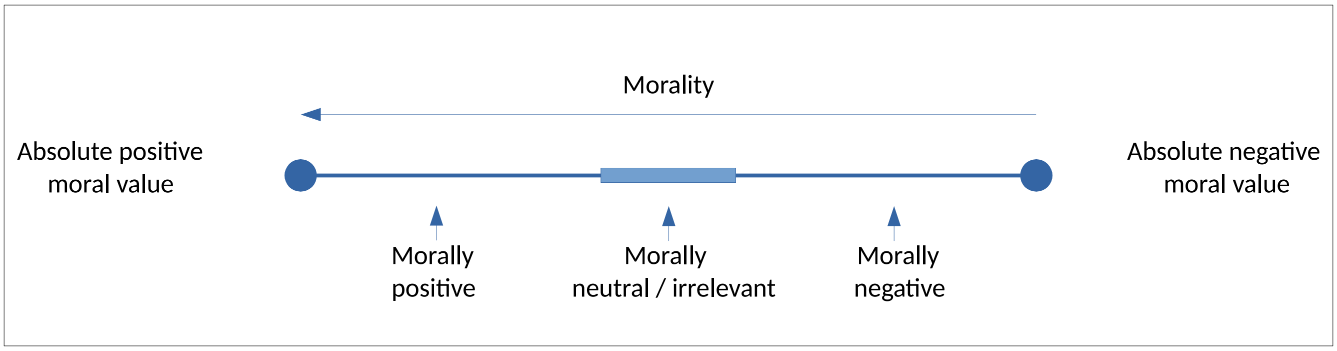
Morality
Absolute positive moral value
Absolute negative moral value
Morally positive
Morally neutral / irrelevant
Morally negative

Illustration 1: The value spectrum

As can be seen from the illustration, morality has the purpose of moving the moral entity, the community, and the world towards the morally positive, and ultimately, towards the absolute positive moral value. Even ethical variants that aim to preserve states of matters as they are (e.g. conservative morality) do this with the belief that this goal represents the absolute positive moral value.

Some ethical theories hold that the aspiration towards absolute positive moral value is of pluralistic nature. For example, virtue ethics typically hold a plethora of virtues that are to be excelled at independently. Thus, for every separate and independent aspiration there would be a separate value spectrum. Whether these can later be taken to form a single, unified one depends on interpretation and is still a heated discussion.

A sample value spectrum can represent states of matters in a general or averaging sense (e.g. as in simple consequentialism); as separate, discrete events judged according to applicable moral norms (e.g. as in deontology); the strengths of virtues, or the general virtuosity of character of a particular person; or another form of morally relevant phenomena.

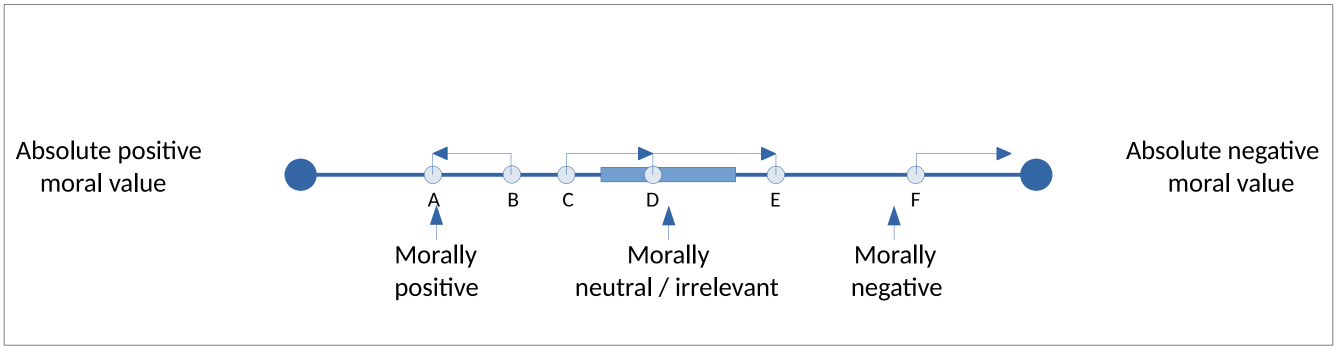
Absolute positive moral value
Absolute negative moral value
A
B
C
D
E
F
Morally positive
Morally neutral / irrelevant
Morally negative

Illustration 2: An example value spectrum

See, for example, Illustration 2. If we take the points (A to F) to represent character virtuosity of separate personae, we can also track their development towards or away the absolute positive moral value at a particular point in, or through, time i.e. from point t_1 to t_2 (not shown). Persona B moved from good to better, reaching character virtuosity level of persona A. Persona C moved to position of persona D, while at the same time D moved to E. And F went significantly towards a moral disaster (and might burn in hell).

If we, instead, have in mind a consequentialist analysis of discrete states of matters, B moved towards the Good and reached point A; while C, D, and F all moved towards evil. We can later average these movements to determine the overall state of matters in a simple consequentialist manner. Or, if we analyze discrete actions according to applicable moral norms, we can say that, for example, a moral entity made action B which was supererogatory, and thus moved the *Rightness* of its actions toward the Right. But it also moved from a supererogatory position C to D, and arrived at the morally neutral or irrelevant. With action D, however, it exited the morally neutral towards the Wrong (i.e. not-Right) by disrespecting an explicit moral norm. And with F it just created a moral catastrophe by stacking several moral norms that it disrespected. And so on and so forth.

The value spectrum could be represented mathematically as a period: [1, -1], where the positive is mathematically positive, and the negative is mathematically negative. Or it may be represented as a period: [0, 1]. This enables mathematical operations, something which will become important later in the metaethics and ethics chapters of this text.

3.1.2 What is of moral value

When discussing value in a moral theory we ought to specify what exactly do we mean that *something* is *valuable*. We need to specify what that something is, how much value does it hold, and what are we to do with it. The term 'value' has seemingly different connotations in different contexts it is used. For example, what is of value (i.e. worthy of pursuit) for a system (e.g. achieving a certain state of matters, a value of a variable, following through a course of action, etc.) might not be taken as the same meaning the term holds when discussing what is of moral value for a person.

But I did include the word 'seemingly' above. The reason is that I will attempt to show in Chapter III. Towards Ethics of Systems (the Metaethics) that moral and systemic value both stem from the same origin—achievement of (personal) goals and conservation of personal continuum; and, from a metaethical perspective, they are tightly intertwined.

Various ethical theories consider various things to be valuable. But they make a typical distinction between *intrinsic* and *instrumental* value (Schroeder, 2016). The end goal of achievement or preservation is that of intrinsic value—the good in itself (as opposed to *good-for-something* instrumental value).

The most commonly mentioned types are hedonistic theories, desire theories, perfectionist theories, comparison and aggregation theories, organic unities theories, and environmental theories (Hurka, 2006; p. 357). Hedonistic ones hold that only pleasure is intrinsically good, and pain is intrinsically bad. Desire theories define desire as what's intrinsically good. These also enable easier identification and measurement of value. Perfectionist theories are focused on excellence in some property/ies (e.g. virtue ethics, but also general theories of purpose). According to these, humans and/or animals and/or other entities have certain properties and purposes in which they (ought to) strive to excel, and this is the ultimate value. Comparison and aggregation theories offer methods to compare and aggregate different values in an attempt to present their value effects locally and/or for the universe. Theories of organic unities define value at different levels, or layers, of unity. They point out that the value held by a unified whole need not be directly comparable (substantially, or in amount) to value of its components. Finally, environmental theories extend value to non-human animals, plants, ecosystems and/or other entities in general. The view I will present later on, the Ethics of Systems view, holds such sentiments towards the entities (systems) of the universe.

Additionally, some ethical theories also recognize a difference between a partial, egoistic good *for a person* (which can also include the separate good for all persons combined i.e. universalizable egoism), and good *for the Universe (God)* (Schroeder, 2016). We will see that this difference in perspective is very important for the work I lay out in the following chapters.

3.1.3 Good as the absence of bad

As I mentioned above, some moral philosophers deny that value can be represented on a spectrum (such as in the section above this one)—see, for example, (Schroeder, 2016; ch. 1.2.1. para. 2) and (Schroeder, 2016; sect. 2.2.3.). This is known as *incommensurability:* the incomparability of distinct state of affairs. It is a more common criticism for deontological theories, where some authors deny the comparability between different kinds of duty, goodness, goodness for, etc. For example, some theorists believe that there is no such thing or state of matters that can be *better* than what is *good* (Schroeder, 2016; sect. 1.2.2.). Another example is the commonly and even intuitively denied consequentialist calculus when dealing with Trolley problems, doctors that save 5 lives at the expense of one, and the like (see section Consequentialism (teleological Ethics), Criticism below).

The reason for this might lay in a different and very important perspective, especially for this work. The reason might be because what is good is not some positive state of matters to be achieved in the world, but, instead, **absence of moral bad!** In this respect, the Good cannot be achieved in a positive sense. Instead, entities can only attempt to protect existence (Being, which is (part of) the Good) from moral bad (see, for example, a patient-oriented view in Floridi's Ethics of Information (Floridi, 2013)). Therefore, it makes no sense to speak of anything better in a positive sense—only in a negative, *less-bad* sense.

Similarly, C. S. Lewis argues that

> "… when we carefully analyze bad and evil we can see that they always represent a striving to achieve something good, but in a wrongful manner. A man can be good because of goodness itself, but he cannot be evil because of evil itself. You can be kind even when you don't feel like being kind, when kindness gives you no pleasure, just because you know that kindness is something right; but, nobody acts in a cruel fashion just because cruelty is something wrong, but because cruelty causes him some satisfaction or it brings him some gain. In other words, evil cannot be successful even in its own evil, while good can be successful even just because the fact that it is good. Good, so to say, exists on its own: evil is just broken good" [translation from Macedonian mine] (Lewis, 2017; p. 69; Macedonian translation).

This subject will be explored in more detail in Chapter III. Towards Ethics of Systems (the Metaethics).

3.2 Argumentation theory

Argumentation theory deals with arguments. But arguments don't exist in a vacuum. The basic requirement for arguments is that there might be an opposition to a claim, a disagreement. Thus, argumentation theory attempts to develop reasoning in a context of disagreement (Liao, 2019). Sometimes arguments are seen as the "basic unit of reasoning", which is the primary consideration of logic (Malerba, 2017; p. 31).

Argumentation theory has been developed to address issues that plague classic logical approaches, by developing contextuality; constrains by procedural rules; reasoning in presence of new information, exceptions and special cases (i.e. provisional validity and defeasiblity). It also helps in "… understanding of what can be defined correct reasoning, studying how reasons support conclusions, what rules regulate the inferential process, how to distinguish a good argument from a bad one in a chain of reasoning, what are the purposes of the reasoning itself (Walton 2005)" (Malerba, 2017; p. 29). There are various types of argumentation theory and distinct sets of logical symbols and operators developed for its representation, developed over the years.

The modern study of abstract argumentation has begun with Phan Minh Dung (1995), but authors were working in the field even before this particular moment. Today, notable authors are Douglas Walton (2009), Antonino Rotolo, Giorgio Bongiovanni, Gerald Postema, Chiara Valentini, Cristiano Castelfranchi, Bartosz Brożek, Bart Verheij, Giovanni Sartor (Bongiovanni et al., 2018), Sanjay Modgil, Henry Prakken (Modgil & Prakken, 2014), Yang Gao, Jeremy Pitt, Régis Riveret et al. (2019), Monica Palmirani, Guido Governatori (Guido Governatori & Sartor, 2005), Trevor Bench-Capon (Bench-Capon, Sartor & others, 2000), Leendert van der Torre, Beishui Liao, Marija Slavkovik (Liao, Slavkovik & van der Torre, 2018), Dov Gabbay, Xavier Parent, John Horty, Ron van der Meyden (Gabbay, Horty, Parent, van der Meyden & van der Torre, 2013), and many more. As we can see, modern argumentation theory has received plenty of significant attention both in volume and in quality of research.

But how is argumentation theory relevant for AI ethics, and specifically for the work here?

Argumentation theory can enable representing and managing multi-stakeholder situations of disagreement. In regards of AI design and utilization this can happen on multiple levels. For example, on a policy level (representatives of the) public, governmental bodies, international organizations, corporations, NGO-s and others can disagree about how exactly should policy regarding AI ethics be formalized. Then, interpretation of a piece of policy can be (attempted to be) resolved by forming the debate in an abstract argumentation scenario. Furthermore, the actual effects on terrain can also be managed by utilizing abstract argumentation (like in figuring out what particular action a smart home system ought to take when faced with conflicting interests of different stakeholders; (Liao et al., 2018)). Sometimes AI entities might enter into disagreements with other AI entities, like what might presumably happen when the skies will be swarmed with delivery and other types of UAVs (drones) that will have to negotiate safe passage.

These are all different levels and situations containing disagreements that may be formally represented by formal argumentation, and attempt to solve them with it. But also, argumentation theory can help directly for this work by providing the means to formally represent moral conflicts, their possible resolution, and the emergence of moral rules.

3.3 Decision theory

Decision theory deals with decision-making, as can be inferred from its label. It is concerned with explication of how reasoning entities (typically spoken about as ‘agents’ by authors in this field) reach conclusions and decisions on how to (not) act in a particular context. It is “… concerned with the reasoning underlying an agent’s choices, … decision theory is as much a theory of beliefs, desires and other relevant attitudes as it is a theory of choice; what matters is how these various attitudes (call them “preference attitudes”) cohere together” (Steele & Stefánsson, 2016). Within it, the ‘orthodox’ approach is what is labeled as *expected utility* (EU) theory. This is an approach mainly developed by von Neumann, Morgenstern, Leonard Savage and Richard Jeffrey.

Typically, decision theory deals with *preferences* over *prospects* (i.e. options). An agent can ‘prefer’ option A over option B in a particular scenario, if A is more ‘choice-worthy’ than B. That means that the agent, attempting to maximize expected utility in the scenario, is ordering the options from maximal (most desired) to minimal and even negative expected[6] utility. To be able to exert preference, EU theorists have posited that there are several axioms that must be observed as to build up coherent formal representation. These axioms are *completeness*, *transitivity*, *independence*, *continuity*, *ordering*, *Sure Thing Principle*, *state neutrality*, *non-atomicity*, *averaging*, *impartiality*, and others (depending on context and researcher). These all help to construct a formal system that can explain the decision-making process of an agent (Steele & Stefánsson, 2016).

[6] Expected, because the agent makes a choice based on the information it has, while it can never fully predict how exactly it would roll out in the future.

In short, decision theory provides a useful framework to gain insight into moral reasoning, since it can be represented as exerting preferences over options in a particular moral scenario. That is why I chose to visit it intermittently where it fits in this work, and in a basic manner.

4 Ethics

According to Simon Blackburn, "Ethics is the study of what is of value in general ..." (Blackburn in (Skorupski & others, 2010)). Essentially, for the mid-term future, AI ethics will be a discipline that will directly concern humans and our communities and societies. We are dealing with AI ethics because of the effects that the widespread introduction of these technologies might have on us, our ecosystems, ways of living, organizations, politics, and of course—our ethics. We *ought* (sic) to pay attention to our basic notions of value (in general and in particular), and what we consider valuable (for more detail on value see 3.1 Value theory above). In any case, the effects of these developments will be so profound that we will have no other choice but to pay attention.

Thus any serious research in AI ethics must start by having a firm understanding of the classical and incoming ethical theories that we, humans, have already devised. Without doing this we might miss very important insights from classical ethics and meta-ethics that can aid us significantly on this road. Another argument for this approach is that humans actually (purport to) follow *human* ethics in their everyday morality—whether that is while resolving conflicts, programming a chat program (Reddy, 2017), designing and using a tool that offers statistical predictions in criminal recidivism (Julia Angwin, 2016), or designing and flying airplanes (Al-Jazzeera, 2019). For now the major designers and users of AI are humans—which means that we are designing and using AI entities with our own ethics in minds (and sometimes with no ethic in mind at all, unfortunately). Therefore, digging dipper into what motivates humans to (not) act in a certain way in particular moral scenarios offers invaluable insight into how we might go on about tackling AI ethics.

4.1 Overview of ethical theories

Traditional ethics starts with the three 'classic' ethical (categories of) theories: deontology, consequentialism and virtue ethics. Besides these, there are some newcomers that have been explored in the near past, especially in the last few decades: ethics of care, environmental ethics (and deep ecology), ethics of information, and now—I anticipate—ethics of systems. An overview of them is given in the further text immediately below.

4.1.1 Traditional ethics

4.1.1.1 Deontology

The word 'deontology' derives from the old Greek 'deon' which means *duty*, and '-ology' which means *study*. Therefore, deontology is the study of duty, in any domain (e.g. ethics and law). Duty naturally leads to moral choices and actions. Thus, we might say that deontology is a (cluster of) normative theory that studies what is morally required, permissible, or forbidden (Alexander & Moore, 2016). It is often defined in *opposition* to consequentialism (Alexander & Moore, 2016), because sometimes it dictates moral choices and actions that do not attempt to maximize the good (or utility), or even go directly against it (McNaughton & Rawling, 2006; p. 424). However, even though deontology and consequentialism are commonly taken as opposite, there are potential bridges (for example, rule consequentialism (McNaughton & Rawling, 2006; p. 428)).

For deontologists, the Right has priority over the Good[7] (Alexander & Moore, 2016). What is the right action is judged in relation to the applicable rules (they are their instantiation in the context), and sometimes an action

[7] The Good, as commonly understood in teleological ethics (consequentialism).

is right even though it is neutral, ignorant to, or even directly opposite to maximization of the Good in the situation.

Instantiation simply means applying an applicable abstract rule in context. For example, the rule that states: "it is forbidden to kill a person" is an abstract rule. However, if someone (e.g. person *x*) actually (attempts to) kill person *y* in a particular room of a house in Florence, then that rule is *instantiated* for that particular context. This instantiation is usually different than the instantiation of the same rule for a murder of *z* by person *u* in an apartment in Timişoara.

Arguably, deontology is one of the favorite 'pets' of programmers and system designers since it is (again, arguably) the simplest to implement. This is because it is based on granular rules that, if applicable (which typically translates to: if triggered), are to be followed regardless of the consequences (but rule application can sometimes be overridden, defeated, undercut, and rebutted; as in argumentation theory). As we will see in virtue ethics and especially consequentialism below, that might sometimes prove morally 'unpalatable'.

Its basic characteristics are: constraints, duties of special relationship, options, and agent-relativity (McNaughton & Rawling, 2006; p. 425).

Constraints are a feature that constrains options of moral choices regardless of whether the moral entity pursued good maximization. For example, a person cannot kill someone to save 5 other persons; or cannot lie, even if that means millions or all other people would be murdered i.e. Kant's deontology. Sometimes there are overriding rules that might defeat such absolutism, but nevertheless these are treated as exceptions.

Duties of special relationship are duties that arise out of special commitment that moral entities have acquired, whether voluntary (i.e. a promise to a friend) or involuntary (i.e. duties to parents, to the community, the society …); explicit (i.e. a signed agreement) or implicit (commitments to a relative). These duties limit our freedom of action, even though when taken voluntarily, they do not represent a violation of freedom. For example, parents who voluntarily decided to have a child are now bound by duty to it (and thus limited in their actions to maybe a smaller set of permissible/acceptable ones), even though that cannot be considered as an infringement upon their freedom of action.

Options are points of the achievement of moral duty after which the moral entities can refuse to do more. They have the option to decline providing more contribution, even though they may be able to provide it, since that might put them in an untenable situation i.e. helping and even sacrificing oneself for the members of one's community is obligatory, but that doesn't directly translate to helping every person in the world, which is out of practical reach for most moral entities.

Agent-relativity means that there are agent-relative moral reasons and constraints. Thus, one can have duty to one's own son before such duty to someone else's child (whereas, in contrast, a simple consequentialism might dictate one's greater duty to care about other people's children while neglecting his own, if that increases the total amount of parental care-giving in the world). Constraints are also agent-relative. For example, if not telling the truth is absolutely forbidden, one cannot lie for the reason that it will stop someone else to lie (or do something else that's forbidden). Both moral entities have the individual, *separate*, personally-applicable duty not to lie.

The main approaches in deontology are the Rossian (W. D. Ross), particularism, the Scanlonian (T. Scanlon), and Kantianism (I. Kant). These all differ on some particularities in regards of what classifies an (im)moral action, or a thing or process of value. However, they all offer an alternative to consequentialism. This alternative can perform better on social relations and autonomy (McNaughton & Rawling, 2006; p. 441). There is an additional split between agent-centered and patient-centered deontological theories. This split will become important as I progress with the text forward. And finally, there is a (contested) class of deontological theories called contractarian. These, however, seem more likely to be meta-ethical than normative, and it is not exactly clear that they would inevitably develop into deontological approaches (as opposed to consequentialist, for example).

Criticism

A common criticism of deontology is around instantiation. Deontologists typically have few words to share about how exactly to know when a certain rule is applicable and ought to be instantiated. What *x* understands as a murder may for *y* well be manslaughter. What x and y understand as a killed person (by another) may mean suicide for *z*, if z holds additional (or sometimes less) information about the case. If there are two general, mutually exclusive rules (1. taking a person's life is forbidden; and 2. suicide is forbidden) and two derived rules from 1. (i.e. 1.1. taking a person's life intentionally is forbidden; and, 2. taking a person's life negligently is forbidden), then, how do we know *which exact* rule to apply (instantiate) in a particular case? Suppose our answer is straightforward: it depends on the set of possible rules and the set of facts about the case. However, we have yet to discover such property of the universe that directly and unquestionably links legal and moral facts to certain, 'applicable' legal and moral rules. What we are left with is (weak) interpretation. Additionally, what do we do when there is an immoral act for which no particular derivative rule yet exists? It would mean that we would need to derive a new rule from a more fundamental one in an ad hoc manner (if possible). However, again, what *exactly* provides support to our derivation of the new ad hoc instantiation of rule *a* over rule *b*, and in that particular form? How exactly to resolve conflicts between rules? The jump from semantics to syntax and vice versa is still a hard problem for us to solve.

There is some work in this direction with argumentation theory, but all the inferences thus set are necessarily defeasible (and therefore weakened).

This also leaves us with the need to derive further, more precise rules that attempt to carry the spirit of the original fundamental rules (or what Ross would define as distinct underivative agent-relative moral considerations (McNaughton & Rawling, 2006; p. 432). But this can quickly end in a moral and legal disaster, where we just derive or invent new rules for every possible situation, clogging the moral and legal traffic (what is known as over-regulation) and making following rules, and thus being morally and legally right, practically impossible.

And even if we do derive or invent a new rule for every possible situation, the very inference from a fact to a rule application is never straightforward, but (for now) always defeasible and weak; because it is commonly performed using commonsense reasoning, and not strong logically valid reasoning.

Another strong criticism is what is known as the 'paradox of deontology' (mainly in patient-oriented theories). Suppose we have persons (A)nne, (B)rad, and (C)harlie. If respecting A's and B's (separate) rights is as important as respecting C's rights, then why is it not permissible or even obligatory to violate C's rights if doing so is necessary to protect A's and B's ones (and hence make an unexpected jump into consequentialism)? (Alexander & Moore, 2016). Thus, paradoxically, following deontological morality might also result in leaving the world in a cumulatively worse moral condition—simply because following the rules is paramount, and not Good maximization. But this is a common problem with all theories that deny comparability (commensurability) or aggregation of doings of wrong, right, or good.

Formal representation and relevance to AI ethics

The main approaches at formal representation and modeling of deontological theories are based on, not surprisingly, deontic logic (i.e. (Gabbay et al., 2013)) and its extension (of a sort), argumentation theory (i.e.

(Liao et al., 2018) (Bongiovanni et al., 2018). I pay more attention to these approaches as I develop my research further down. They are obviously valuable when dealing with AI ethics since they provide a well-formed formal representation of reasoning, decision-making, and action that can be designed as a cybernetic part[8] of an AI entity.

4.1.1.2 Consequentialism (teleological Ethics)

One criticism of the so-called 'smart contracts' is that they are, actually, very dumb and typically don't take context into account. They simply follow rules—even when these rules might result in dire consequences. This 'blind rule following' is one of the consistent critiques also of deontology in general (see above).

This is where teleological ethics comes into play. The word is derived from 'telos' and '-ology', which means study of goals and consequences (or what programmers and system designers might sometimes refer to as 'utility maximization').

Teleological (consequentialist) ethical theories put the Good before the Right (Alexander & Moore, 2016) (Brink, 2006; p. 381) (Robertson, 2006; p. 440), but what this Good is can vary, and can take a monist or a pluralist form. In general, moving towards the Good is understood as maximization of value by making choices. According to consequentialists, morally right choices are those that increase the Good (as defined in the particular consequentialist moral theory)—directly or indirectly, immediately or in the long run, on average or cumulatively, or in general (McNaughton & Rawling, 2006; p. 428 - 431).

Furthermore, consequentialist theories are typically *agent-neutral* (compare with agent-relativity in *Deontology* above). This means that "… valuable states of affairs are states of affairs that all agents have reason to achieve without regard to whether such states of affairs are achieved through the exercise of one's agency or not" (Alexander & Moore, 2016). However, there are consequentialist theories that try to accommodate agent-relativity, such as *self-referential altruism* and *ethical consequentialist egoism* (Brink, 2006; p. 382).

There are different flavors of consequentialism (Brink, 2006; p. 381 - 384) (McNaughton & Rawling, 2006; p. 428 - 431). These bring about different conclusions about what a moral entity ought to do in a particular moral scenario and in general. Therefore, a generalized analysis of consequentialism is a difficult endeavor. Under direct consequentialism (DC) we have act consequentialism (AC) and scalar consequentialism (SC). AC seeks to maximize value directly, where the entity "… should perform that action whose value (of the relevant sort) is at least as great as that of any alternative available to her (or at least one such action, if there are multiple actions meeting this condition)" (Brink, 2006; p. 383). In short, the moral entity ought to choose the action out of all possible ones that holds the greatest value. SC permits the moral entity to choose an action that solely increases the Good, rather than maximizing it in the scenario. Thus, an action that is good *enough* (i.e. passes a certain threshold) can be chosen, and not necessarily the 'best' one. Rule consequentialism (RC) is part of indirect consequentialism (InDC) and focuses on good *rules* instead of good actions. The goodness of actions is judged by the goodness of the rules under which the actions can be subsumed. Similarly, motive consequentialism (MC) is also part of InDC, but instead on focusing on good rules it focuses on good *motives*. Actions that can be subsumed under good motives are good—and vice versa. And finally, we have sophisticated consequentialism (SopC), also within InDC, where the moral entity seeks to lead an objectively consequentialist life, but not necessarily subjectively consequentialist one (McNaughton & Rawling, 2006; p. 429).

[8] Notice that I do not use the word 'component', as such 'part' might even be the whole entity itself, or some of its dispositions to act and reason according to context—even without having a *special*, *separate* component dedicated for this.

Criticism

Consequentialism is not without its critics, though. Two critiques are pretty successfully aimed at it, for which what is typically and widely understood as consequentialism has no good defense. Those are that consequentialism is either *too* demanding, and in parallel (and perhaps ironically), that it is *not demanding enough* (Alexander & Moore, 2016).

The first critique is that consequentialism requires *too much* from moral entities. For consequentialists, there simply are no actions that are morally irrelevant. Every thing an entity in the world does (or does not) is either forbidden or required. This leaves no space for supererogation or for simple moral neutrality—two notions that are intuitive to us, but 'classic' consequentialism cannot account for in a satisfactory manner. The second critique is that consequentialism *does not require enough*. In most consequentialist theories there is no space for special preference (partiality) for personal projects, close persons, countrymen, friends, family, organizations and the like (notions that are, again, intuitive to our moral senses). In short, consequentialism typically does not follow any kind of locality principle, because it might result in partiality.

But the above are not the only strong critiques of consequentialism. For example, simple teleological ethics (i.e. maximize the Good without regards to anything else) can lead to permission to kill, rape, pillage, abandon, lie, and deprive—but only if it more beneficial than harmful (Alexander & Moore, 2016). Thus, a doctor may be permitted to kill a healthy person in order to save 5 dying ones with his or her organs. What is particularly interesting for AI ethics is the modification of the trolley problem—where a fat man can be pushed on the tracks, thus saving several lives at the expense of his own. According to simple consequentialism this would be permitted, even though many (if not most) people's moral intuitions would strongly be repelled by the idea.

Consequentialists, of course, have answers to these critiques (Alexander & Moore, 2016). Some argue that it is 'enough' that only a certain (and not the total possible) level of Good is achieved (which is known as 'satisficing', sometimes promoted as *scalar consequentialism* (Brink, 2006; p. 383)). Others introduce the distinction between positive and negative duties; where negative duty is not to make the world worse (*non-maleficence*), where a positive duty is to make it better (beneficence; this distinction also permeates Floridi's ethics of information (IE) and its 4 ethical principles; (Floridi, 2013; p. 71). Non-maleficence and beneficence are not directly connected. Thus, saving 5 dying people cannot lead to killing a single healthy person; and a fat person is not permitted to be pushed to save others on the trolley track (this is similar to constraints in deontology, as discussed above). A third answer is a move from maximizing the Good to maximizing good *rules* as a primary target (and indirectly assessing actions according to those rules). This is known as *rule consequentialism* (RC). Intuitively, RC reminds us a lot of deontology, which is why it serves as a potential bridge between the two categories of ethical theories (as I mentioned above in *4.1.1.1 Deontology*).

Relevance to AI ethics

As with the other 'classic' ethical theories, consequentialism (teleological ethics) is of both general and specific relevance for AI ethics. Of general, since it is one of the dominant ethical theories of humans. If AI is to follow human moral intuitions and reasoning, it should be able to accommodate teleological ethics where appropriate and expected. But consequentialism can also be used as a guidance to program algorithms that, for example, attempt to maximize or satisfice a certain variable (i.e. utility maximization; like in trading algorithms that attempt to maximize profits, or distribution algorithms that attempt to achieve the best resource distribution possible).

4.1.1.3 VIRTUE ETHICS

Virtue ethics is a category term that includes ethical theories focused on character development, and the centrality of virtues for developing and living the good life (Athanassoulis, 2019). This is contrasted to following one's duty (deontology) or acting to achieve best consequences (consequentialism). Virtue ethics is primarily concerned with questions like: "How should I live?" and "What is the good life?"; instead of trying to devise a set of universal principles to apply in context.

For virtue ethics, there are three concepts that are central: practical wisdom, virtues, and eudaimonia (Hursthouse & Pettigrove, 2018).

Practical wisdom holds a central role in virtue ethics since how a person ought to act in a certain situation is contextual on the person itself, and should be thought of and tailored appropriately by himself. Practical wisdom is what enables a person to discover for himself the best way to achieve virtue i.e. to develop the various virtues, which in turn will enable him to live a virtuous life and thus flourish. But in order to achieve this, a person first needs to become experienced in how both the external and internal worlds work. This is typically developed through several avenues: trial and error, education (including moral one), tradition and established rules, and scientific study. Only after a certain amount of experience a person can be expected to have extracted enough practical wisdom to be able to further his own development into more virtuous living.

Virtues, unsurprisingly, also take a central role in (most) virtue ethics, since they are conscious dispositions to act in a certain manner (excellent character traits) that fit living the good life. In virtue ethics acting virtuously is bound on being internally virtuous. A young, inexperienced person can (naïvely) act in what seems like a virtuous manner (for example, if they were told to act in such manner by their parents or by already established rules in the community); but they will not be virtuous, since they have not developed the character trait internally that would result in such actions. Thus, for virtue ethicists, one can only be characterized as an honest or courageous person if they are truly and consciously honest and courageous internally—not only if they simply act like it (Hursthouse & Pettigrove, 2018) (Annas, 2006; p. 517). Roots of virtues may be natural parts of character, but a person has to develop them to excellence and integrate them in his person by using practical wisdom, conscious effort, and self- and world-discovery (as mentioned above). (Self-)consciousness is very important, since a virtue is never a habit, but always a conscious effort (Annas, 2006; p. 516).

Virtue ethics is concerned with moral entities (directly with agents, indirectly with patients); in contrast to consequentialism and deontology, who are concerned with (in)actions regarding their effects/states of matters, or rules and duties. Thus, it is a refreshing view that brings back the focus on the entities themselves. Consequentialism (some forms more than others) can sometimes be attacked with the reason that it disregards moral entities as individuals, and just ignorantly aggregates them. Whereas, virtue ethics is concerned with flourishing of the individual; and living virtuously is purported to improve flourishing of his environment.

To cite from Annas (2006; p. 517):

> "The virtuous agent, then, does the right thing, undividedly, for the right reason—he understands, that is, that this is the right thing to do.
>
> …
>
> For virtue ethics, the purpose of good moral education is to get the pupil to think for himself about the reasons on which he acts, and so the content of what he has been taught. Ideally, then, the learner will begin to reflect for himself on what he has accepted, will detect and deal with inconsistencies, and will try to make his judgments and practice coherent in terms of a wider understanding which enables him to unify, explain and justify the particular decisions he makes. This is a process that requires the agent

> at every stage to use his mind, to think about what he is doing and to try to achieve understanding of it (...)".

The end 'goal' of virtuosity is to live the good life i.e. to flourish. By excelling in the virtues, a person will live his life as a whole in a way that is valuable to live. This end is often called *eudaimonia*, with its currently accepted best translation as *flourishing*.

In order for a person to get on the path of virtuosity, it is well advised to start with already established rules, education, advice, and role models. To try to determine why these represent living virtuously, how they fit or differ his own context and character, and then develop virtue organically that naturally follows into virtuous action.

Criticism

A critique of virtue ethics in regards of AI is that AI entities are not (self)conscious (for now, at least), and thus this approach makes no sense for them. There is no way an AI entity can make self-reflection, and attempt to excel in (appropriate) virtues, so that this results in it becoming virtuous. This might appear as a strong argument at first glance. However, there are some considerations that weaken it.

Firstly, as we will see further down, consciousness is an intrinsic part of any system's existence (i.e. integrated information; (Tononi et al., 2016)) and functioning (although consciousness does not directly translate to self-consciousness, with which virtue ethics is arguably concerned; but the jump between consciousness and self-consciousness is not a 'hard' one). Furthermore, even if a system is not self-conscious, it can act *as if* it is (Annas, 2006; p. 528). That is to say, system designers can design a system to act as would a self-conscious virtuous entity act in such a situation. We can sample virtuous humans in real or hypothetical situations and use this as a benchmark to test the algorithms (something we are actually doing even today, mostly implicitly). Even more, virtues and virtue calculus might even be encoded inside the system. That won't necessarily mean that such systems will truly become virtuous, but for all practical external effects—they will be.

Another criticism might be that virtue ethics is aimed at the individual living a virtuous life and thus flourishing. This might be seen as an egoistic, selfish approach (Annas, 2006; p. 530), unfit for dealing with morality and ethics. However, even the classical and ancient virtue theorists have stressed that living virtuously as an individual will result in flourishing, not only of the individual, but also of the environment around him. The 'selfish' criticism is not unique for virtue ethics. In ethics, in any case, we have *separate* moral entities that have to decide on their next moral choice. Even if they decide to sacrifice themselves to help others, this still might seem selfish, as they do it to satisfy their own internal goals and ideals. But the *effects* of their actions are acutely not selfish—on the contrary. Hence, similarly with virtue ethics.

A third criticism is that virtues are not sufficient, or even necessary, for flourishing. This critique is trying to posit that it is a (instrumental or substantial) mistake, or implausibility, to assume that developing virtues will lead to personal and environmental flourishing. It is a strong critique around which a lively debate is running currently, so there is not a definite refutation or support for it.

Relevance to AI ethics

In the short term we cannot expect for AI systems to become truly virtuous, since they lack self-awareness and the complexity of internal characters that people have. This might seem as an eliminating factor for virtue ethics out of AI ethics. However, as mentioned above, we can attempt to develop algorithms and systems that act *as if* they are virtuous.

To do this we would need to study models of virtuous persons, and how they will go about making decisions in situations where artificial systems will be involved. That is, to use human virtuous role models[9]. For example, if an algorithm is to be employed to help judicial decision-making, the best course of action might be to study how the best (most virtuous) human judges make their judgments—and then try to transfer this into the AI system itself.

Other examples can include watching virtuous persons deliberate and decide on trolley problems, trading, exchange, security, privacy, trust … and attempt to emulate this within AI entities. Should that autonomous vehicle kill the dog on the left, the little boy on the right, or the passenger inside? Ask (hundreds and thousands of) virtuous persons and see what they would answer (as was done with (assumed) virtuous persons on the Internet by the MIT's The Moral Machine experiment; (Awad et al., 2018)).

Let's also not forget the virtuosity required of AI designers and programmers. Before AI entities have the capacity to reason ethically and adapt to moral requirements, their reasoning capacities will be designed and implemented by humans. To avoid using these systems for nefarious and immoral/unethical purposes, the designers and implementers themselves need to be striving for the good, and avoid the bad; to live virtuously and thus *program and design* virtuously.

4.1.2 Newcomers

4.1.2.1 Ethics of information

This was already covered in section 2.2 Philosophy of information and especially 2.2.1 Ethics of information.

4.1.2.2 Ethics of systems

This is a potential multidisciplinary area of study that attempts to make a bridge between systems science and ethics, similarly to how a bridge was built between information science and ethics—with ethics of information. It is what I attempt to perform here in my work, and as largely an upcoming field, there is not a wealth of research and publications on the subject.

However, there is some work already being done. For example, the *International Journal of Ethics and Systems* explicitly states that the focus of the Journal

> … is on disseminating the theory and practice of morality and ethics as a system-oriented study defined by inter-causality between critical variables of given problems. (Emerald Publishing, 2019)

For now, though, this journal seems more focused on economical analysis, which is only one part of the multidisciplinary approach required by this field.

Sometimes, there are authors that come close to the field by discussing, for example, the ethics of systems thinking (Harter, Dean & Evanecky, 2004), embedded ethics in systems (Key, Azab & Clark, 2019) (Bonnemains, Saurel & Tessier, 2018) (Thekkilakattil & Dodig-Crnkovic, 2015), using general systems theory for information systems research (Kanungo & Jain, 2007), ethics of holism (Martin, 2014), and similar.

However, it is obvious that the level of research in this potential field is still weak, scattered, and just starting. What's more, it is typically not aimed at direct application of systems science to ethics and vice versa (although there are exceptions; for example: Nuotio (2010)). Therefore, with this work I intend to contribute to the birth of the field, which I dubbed the *Ethics of Systems*.

[9] Akin to how game production companies use real humans to determine how the human body moves, so that they can program movements of non-player characters that seem natural.

4.1.2.3 Environmental ethics and deep ecology

Environmental ethics (EE) is concerned with the moral status and value of the environment, ecosystems, and all non-human parts of nature. It is also concerned with the ethical relationships of human beings with the environment (Cochrane, 2006). Of course, by extension, it can be concerned with all non-human parts of the universe at large such as natural objects, processes and systems (Brennan & Lo, 2010; p. 754); although I am inclined to argue that other disciplines such as *ethics of information* or *ethics of systems* are better suited to handle this scope.

EE is an approach challenging anthropocentrism typical of human ethical discourse (Cochrane, 2006). It is also a patient-oriented ethics (Floridi, 2013; p. 63). It attempts to answer two questions: 1. what is the moral value and moral status of the environment and its non-human elements? (Brennan & Lo, 2016); and, 2. why is this so? By extension, a question on what moral relationship humans ought to have with the environment can also be studied (Cochrane, 2006) (Brennan & Lo, 2016).

This last question is an extension because it is an attempt to challenge anthropocentrism, as already mentioned above. For example, the Australian philosopher Richard Routley describes and criticizes the notion of "human chauvinism", which is, according to him, the mainline principle of western liberal thinkers. He goes on to say that: "Whether the blue whale survives, [...] should not have to depend on what humans know or what they see on television. Human interests and preferences are far too parochial to provide a satisfactory basis for deciding on what is environmentally desirable" (Routley, 1973:210; in Brennan and Lo (2010; p. 755).

EE philosophers typically argue that the environment and its non-human parts have moral value and thus deserve moral status and moral respect *on their own*. That is to say, they do not simply hold instrumental moral value for humans; but hold intrinsic moral value instead. Therefore—animals, species, ecosystems, trees (and for some theories even holistic entities such as mountains, rivers, and even planets and solar systems; and sometimes at odds with each other; Cochrane (2006))—can be taken as intrinsically valuable, regardless if humans recognize this or not. Some of the concerns of EE are the preservation of biodiversity, sustainability, climate change, pollution, exploitation of natural resources, recycling, renewable energy, and similar.

Renowned authors in this field are Peter Singer, Aldo Leopold, Tom Regan, Robin Attfield, John Benson, Murray Bookchin, Michael Boylan, Rachel Carson, Joseph R. DesJardins, Warwick Fox, Lawrence E. Johnson, Mark Sagoff, Arne Næss, Sigmund Kvaløy, Nils Faarlund, Andrew Brennan, and others (Cochrane, 2006) (Brennan & Lo, 2016) (Brennan & Lo, 2010).

Deep ecology movement

EE is also naturally connected with the *deep ecology movement*, started in Scandinavia in the 70's by Arne Næss and his colleagues Sigmund Kvaløy and Nils Faarlund (Brennan & Lo, 2016). As Næss himself would say, the contrasted *shallow ecology movement* is "'fight against pollution and resource depletion', the central objective of which is 'the health and affluence of people in the developed countries'" (Brennan & Lo, 2016). Thus, *shallow ecology movement* in his view is still irreparably anthropocentric.

In contrast, the *deep ecology movement*: "… endorses 'biospheric egalitarianism', the view that all living things are alike in having value in their own right, independent of their usefulness to others. The deep ecologist respects this intrinsic value, taking care, for example, when walking on the mountainside not to cause unnecessary damage to the plants" (Brennan & Lo, 2016).

Relevance to AI ethics

A comprehensive ethical theory applicable to AI cannot avoid tackling the ethical challenges that the environment poses. AI entities already help to work the land, process food, guide oil tankers (that occasionally spill their oil at the sea), suggest which products (with which plastic types) to buy or deliver, fly airplanes (that emit CO2), map out forests to plant or cut, and a plethora of other processes that affect the environment. It is needless to say that AI entities and systems, if they are to act ethically, ought to take into account the environment, animals, and whole ecosystems—not just humans.

As we can see, both EE and the *deep ecology movement* are ethical[10] approaches that recognize intrinsic moral status (i.e. moral dignity and value) of the environment and all its parts, and thus argue that people ought to have moral respect for it. This is very similar to how Floridi's *ethics of information* treats the environment, even though the ontological and epistemological approach is coming from a different perspective i.e. the world as the infosphere, and the method of levels of abstraction (see Floridi (2013; p. 18) for more). And *ethics of systems* provides even lower level, and potentially even more comprehensive, view on ethical considerations regarding the universe, incorporating *ethics of information*, *environmental ethics*, and *deep ecology* (or at least being compatible with them).

4.1.2.4 Ethics of care (feminine ethics)

Ethics of care (sometimes known as feminine ethics[11]) is an approach in ethics that prioritizes relationships between moral entities (caregivers and caretakers), and the well-being of those moral entities. It "implies that there is moral significance in the fundamental elements of relationships and dependencies in human life" (Sander-Staudt, 2011).

If we go back to the distinction between moral entities (moral agents and moral patients), we can say that ethics of care is patient-oriented ethics. That is because, according to Floridi, patient-oriented ethics (such as bioethics, environmental ethics, and medical ethics), hold the

> "broad view that any form of life has some essential proprieties or moral interests that deserve and demand to be respected, at least initially, minimally, and overridably. They argue that the nature and well-being of the patient of any action constitute (at least partly) its moral standing, and that the latter makes important claims on the interacting agent that, in principle and when possible, ought to contribute to the guidance of the agent's ethical decisions and the constraint of the agent's moral behaviour" (Floridi, 2013; p. 63).

For the various types of ethics of care, the main focus is "the compelling moral salience of attending to and meeting the needs of the particular others for whom we take responsibility", and it "stresses the moral force of the responsibility to respond to the needs of the dependent" (Held, 2006; p. 538). Ethics of care offers also a fresh input which values emotions, instead of denying or avoiding them as important elements of ethics (as is common with the rationalists approaches)—especially since care may be taken as an emotionally-driven disposition, at least typically in humans and other mammals (but not necessarily). It also offers an inverse perspective in regard of the dominant moral theories. Instead of claiming that the more impartial the approach, the better; it rejects that view and claims that partial relationships (like friendships and family) are exactly what needs to be preserved and even improved (Held, 2006; p. 540).

It is easy to see that a care ethics can readily make the jump for care towards all dependent aspects of our environment (akin to the *shallow ecology movement* that I discussed above); then to all these that are in need for care (akin to the *deep ecology movement*); and towards universal care.

[10] 'Ethical' in the sense of falling under the category of ethics, not as in morally and/or ethically positive.

[11] And sometimes, rather erroneously, as feminist ethics.

Relevance to AI ethics

This last aspect (universal care) will be very important especially because it is, not only compatible, but essential part of *ethics of systems*. AI entities will simply have to take care about the parts of other systems that they are utilizing, manipulating, and affecting, or even simply those that need care. For example, an AI entity inside a care robot that does not provide that care would not only be useless, but might also be dangerous.

But why care is so important for AI entities? Taking into consideration the fragility of Being that can be hurt, and of relationships that can be changed for the worse, care can be exactly that morally-guided behavior that maintains them. A useful analogy here would be that as humans (are expected to) take care about those that brought them into life and flourishing (their parents, their community, their society, the world), so will AI entities be expected to care in the same manner.

5 Ethics of AI

This section will certainly be the crux of the reading performed for this study. Published material on AI ethics is far and wide. In the past decade there has been an explosion of articles, books, teaching classes, presentations, conferences, and talks on the subject.

I have initiated the study by going over some recommended books:

- *Robot Ethics: The Ethical and Social Implications of Robotics* (Lin, Abney & Bekey, 2011) with editors Patrick Lin, George A. Bekey, and Keith Abney;
- *A Legal Theory for Autonomous Artificial Agents* (Chopra & White, 2011) by Samir Chopra and Laurence F. White, although focused on legal matters offers a valuable ethical perspective;
- *Superintelligence: Paths*, *Dangers*, *Strategies* (Bostrom, 2014) by Nick Bostrom; and,
- Floridi's *Ethics of Information* (Floridi, 2013).

Cumulatively, they gave me a good overview in the various subdomains that this field touches upon.

For example, Bostrom's *Superintelligence* dives into a plethora of problems expected to arise before and after a purported superintelligent AI enters the world, probably created in a thinly-constrained arms race driven by a zero-sum-like thinking. One of the main problems with this superintelligent entity would be the so-called *control problem*, where humans would probably struggle to control such a versatile intelligent with even with all the necessary precautions in place. We would also probably want such an entity to acquire values (i.e. a value theory of its own) that would at least positively consider human interests and endeavors. This is a problem on its own explored in the book. Bostrom also goes on to explain the assumed *instrumental goals* (sic) of such an entity: self-preservation (sic), goal-content integrity (sic), cognitive enhancement, technological perfection, and resource acquisition (sic) (notice how these emphasized three fit in the *Ethics of Systems* Framework that I work on in Chapter III. Towards Ethics of Systems (the Metaethics)).

Lin, Abney and Bekey's *Robot Ethics* is a formidable volume that gathers several renowned authors in the field to explore issues in ethics, design and programming, law, war, psychology and sex, medicine and care, and rights and responsibilities. The book itself has an introductory part where the basics of ethics and AI ethics are explored, and then goes on in deeper waters. Some of the subjects that are explored are robotic (or rather: artificial) personhood; the problem of intractability; moral machines; a Buddhist and Divine Command approach to AI ethics; responsibility for military AI entities; legal personhood; robotic lovers and caregivers; the thread of ethical nihilism that may be brought by (in)ethical AI entities; and many more.

Chopra and White's *A Legal Theory for Autonomous Artificial Agents* dives deeply into artificial agency and all the potential issues that spring out of it: contracts, attribution of knowledge, tort liability, and personhood. Both authors make this formidable, and pretty successful, effort to form a complete legal theory revolving around the aforementioned. The book focuses on using 'agents' as the guiding factor, where the authors define agents as "Intuitively, an agent is something able to take actions. One way to distinguish agents from other entities is that agents do things, as opposed to have things happen to them; to deny something or someone agency is to deny the capacity to take actions, for the actions of the agent distinguish it from the rest of the world" (Chopra & White, 2011; p. 11). Unfortunately, it doesn't directly explore the same legal theory on the other side—the side of patients. But there is plenty of other material to work on this side, as I already mentioned.

And I took an in-depth overview on Floridi's *Ethics of Information* earlier in 2.2.1 Ethics of information.

From here, I made a wide search through publication databases (e.g. Elsevier, Routledge, Oxford University Press, Cambridge University Press, Google Scholar, the Semantic Scholar, etc.) to find the most important and recent publications on the subject in the article form. This resulted in a plethora of documents from which I built up my database initially, and on which I build up onward as I discovered new material. My database on AI ethics articles, books and proceedings is currently at above 286 entries, which is more than 1/5 of all the relevant documents I obtained for this study. Combined with more than 277 documents for ethics in general, and more than 123 documents for AI and robotics in general, they comprise around 3/5 of whole database of 1,195 documents for the study.

A general impression that can be extracted out of the variety of documents I came across is that it is common for researchers coming from technical fields (e.g. AI programming) to not be very strong on the ethics side; and vice versa, for ethicists and philosophers not to be strong in the technical side of matters. Following the two facets of the issue in significant depth is crucial if we are to perform good research and offer valid conclusions, advice, and opinions. Admittedly, this is very hard to do; but there are scientific research programs (such as LAST-JD, on which I was accepted to do this study) that aim to do exactly that—attempt to form a multidisciplinary bridge between the different aspects of the same problematic.

There is a plethora of themes and subjects that are dealt in the AI ethics publishing:

- privacy (Floridi et al., 2018; p. 11) (AIHLEG, 2019) (Crawford et al., 2016) (Ambrose, 2014) (Ananny & Crawford, 2016) (Bechor, Zhang & Cruz, 2018) (Brundage et al., 2018) (Campolo, Sanfilippo, Whittaker, Crawford & Selbst, 2017) (Collingwood, 2018) (Danaher et al., 2017) (Delvaux, 2016);
- consciousness in AI entities, and its relation to ethics (DiCarlo, 2016) (Bello, Licato & Bringsjord, 2015) (Torrance, 2008) (Torrance, 2014), agency (Himma, 2009), and patiency (Gunkel & Bryson, 2014);
- information and computer ethics (Einar Himma, 2007) (Floridi, 2013) (Floridi, 2010);
- Good AI society (i.e. society which becomes morally good with the use of AI) and AI as a force for good (Cath, Wachter, Mittelstadt, Taddeo & Floridi, 2017) (Taddeo & Floridi, 2018) (Floridi et al., 2018);
- transparency, explainability, and explicability (Wachter, Mittelstadt & Floridi, 2017);
- AI and robotic accountability (Caplan, Donovan, Hanson & Matthews, 2018) (Martin, 2018) (Kahn Jr et al., 2012) (Wachter et al., 2017) (Ananny & Crawford, 2016) (Martin, 2018) (Kahn Jr et al., 2012);
- moral decision-making, formal modeling and programming moral reasoning in AI (Conitzer, Sinnott-Armstrong, Borg, Deng & Kramer, 2017) (Wallach, 2010) (Wallach, Franklin & Allen, 2010) (Pereira,

Saptawijaya & others, 2016) (Saptawijaya, 2015) (Bringsjord, Arkoudas & Bello, 2006) (Liao et al., 2018) (Goodall, 2017) (Bringsjord et al., 2006) (Criado, Argente, Noriega & Botti, 2013);

- the control problem (Kleeman, 2017) (Sullins, 2013) (Bostrom, 2014) (Bello et al., 2015);
- taxonomy (Franklin & Graesser, 1996) ;
- AI weapons (Meizhen & Zhaoming, 2016) (Fleischman, 2015) (Sullins, 2013) (Schulzke, 2011) and distributed responsibility (Schulzke, 2013);
- autonomous vehicles and their decision making (Schäffner, 2018) (Awad et al., 2018) (Collingwood, 2018) (Goodall, 2014a) (Goodall, 2014b);
- risk management (Goodall, 2016);
- fairness (Patrignani & Whitehouse, 2015) (Patrignani & Whitehouse, 2014);
- opacity in AI decision-making (Burrell, 2016);
- ethical, social, and economic implications and challenges of AI now and in the future (Crawford et al., 2016) (Stankovic, Gupta, Rossert, Myers & Nicoli, 2017) (Martin, 2018) (Meek, Barham, Beltaif, Kaadoor & Akhter, 2016) (Bechor et al., 2018) (Buechner, 2018) (Muehlhauser & Helm, 2012);
- algorithmic governance (Danaher, 2016) (Danaher, 2015) (Danaher et al., 2017);
- AI and trust (AIHLEG, 2019) (Collingwood, 2018) (Tavani, 2015) (Grodzinsky, Wolf & Miller, 2011) (Lim, Stocker & Larkin, 2008);
- ethical design, engineering and deployment of AI (Kitto & Sylvester, 2002) (Grodzinsky, Miller. & Wolf, 2008) (Arnold & Scheutz, 2016);
- robot and AI rights (Tavani, 2018) (Gunkel, 2014) (Ashrafian, 2015a) (Richardson, 2016) (Ashrafian, 2015b) (Gunkel, 2017);
- moral philosophy regarding AI (Moor, 2006) (Ashrafian, 2015a) (Gunkel, 2012) (Gunkel, 2014) (Scheutz, 2017);
- moral and ethical frameworks and guidelines (Conitzer et al., 2017) (Dameski, 2018) (Floridi et al., 2018) (Mansouri, Goher & Hosseini, 2017) (Schaerer, Kelley & Nicolescu, 2009) (Wiltshire, 2015) (AIHLEG, 2019);
- and many, many more subjects.

A few additional mentions in the form of books and PhD theses are *Machine Ethics* by editors Michael Anderson and Susan Leigh Anderson (Anderson & Anderson, 2011); Peter Danelson's *Artificial morality: Virtuous robots for virtual games* (Danielson, 2002); Peter Han's *Towards a superintelligent notion of the good: Metaethical considerations on rationality and the good, with the singularity in mind* (Han, 2015); Luís Moniz Pereira and Ari Saptawijaya's *Programming Machine Ethics* (Pereira et al., 2016), as well as Saptawijaya's own PhD thesis work, *Machine ethics via logic programming* (Saptawijaya, 2015); Spyros Tzafestas' *Roboethics: A Navigating Overview* (Tzafestas, 2016); Vincent Muller's *Fundamental Issues of Artificial Intelligence* (Müller, 2016), especially the fifth chapter; and Andrighetto, Governatori, Noriega, van der Torre's *Normative Multi-Agent Systems* (Andrighetto, Governatori, Noriega & van der Torre, 2013).

6 Law

6.1 International and regional documents

6.1.1 Human Rights

Why include human rights in this study? The answer is very simple. Codified human rights are a ethico-legal instrument. That is, they are ethical principles and guidelines that were codified and thus turned into law.

By today, human rights documents established within the framework of the United Nations represent the most authoritative documents on the subject. Other influential documents are the constitutions of countries in the world that also stipulate protection and improvement in respect for human rights.

In any case, countries-members of the United Nations have to sign the Charter of the United Nations, which *de facto* implies signing and implementing the International Bill of Human Rights (IBHR; see below).

The Charter itself contains this text in Article 1:

> **Article 1**
>
> The Purposes of the United Nations are:
>
> 1. To maintain international peace and security, and to that end: to take effective collective measures for the prevention and removal of threats to the peace, and for the suppression of acts of aggression or other breaches of the peace, and to bring about by peaceful means, and in conformity with the principles of justice and international law, adjustment or settlement of international disputes or situations which might lead to a breach of the peace;
>
> 2. To develop friendly relations among nations based on respect for the principle of **equal rights and self-determination of peoples**, and to take other appropriate measures to strengthen universal peace;
>
> 3. To achieve international co-operation in solving international problems of an economic, social, cultural, or humanitarian character, and in **promoting and encouraging respect for human rights and for fundamental freedoms for all** without distinction as to race, sex, language, or religion; and
>
> 4. To be a centre for harmonizing the actions of nations in the attainment of these common ends.
>
> [boldtype mine; Dameski] (United Nations, 1945).

As we can see, support and respect for human rights and freedoms, both individual and collective, are baked right into the United Nations since its inception during World War II. Additionally, they are applicable law in the member-states—signatories of the treaties in the IBHR.

Therefore, human rights have to be taken in account when designing, deploying, and utilizing AI in human civilization. AI entities (and of course their designers and users) have to respect human rights (implicitly or explicitly), and states need to take measures to ensure that respect for human rights is established on their territory in a substantial manner (not just formally).

6.1.1.1 The International Bill of Human Rights

The IBHR represents the very fundamental, and universal, collection of documents that apply universally in this sense to all of humanity.

It contains the Universal Declaration of Human Rights (UDHR), the International Covenant on Civil and Political Rights (ICCPR), the two optional protocols to the ICCPR, and the International Covenant on Economic, Social and Cultural Rights (ICESC).

I will not be going into in-depth exploration of the documents. The purpose here is to get acquainted with an overview of the UN human rights framework. The documents are written in such a way to be commonly understandable by every person in the world that has acquired common sense. For example, the preamble of the UDHR contains the following sentence: "Whereas a common understanding of these rights and freedoms is of the greatest importance for the full realization of this pledge, ..." (United Nations, 1948).

Therefore, I will assume that common-sense understanding of the terms and provisions used in the documents to prescribe the human rights and dignities is correct and enough for the most cases; and will seek to define or clarify them only where needed.

The three categories

There are 3 broad categories by which the human rights, dignities, and related prohibitions are divided (Nowak, Klok, Schwarz, Arbour & Johnsson, 2005; p. 2). These are as follows in Table 2: Categories of human rights, dignities, and prohibitions (with the rights that belong to each category):

Table 2: Categories of human rights, dignities, and prohibitions

Category	Civil and political rights	Economic, social, and cultural rights	Collective rights
Rights, dignities, and / or prohibitions	• Right to life • Freedom from torture and cruel, inhuman or degrading treatment or punishment • Freedom from slavery, servitude and forced labor • Right to liberty and security of person • Right of detained persons to be treated with humanity • Freedom of movement • Right to a fair trial • Prohibition of retroactive criminal laws • Right to recognition as a person before the law • Right to privacy • Freedom of thought, conscience and religion • Freedom of opinion and expression • Prohibition of propaganda for war and of incitement to national, racial or religious hatred • Freedom of assembly • Freedom of association • Right to marry and found a family • Right to take part in the conduct of public affairs, vote, be elected	• Right to work • Right to just and favorable conditions of work • Right to form and join trade unions • Right to social security • Protection of the family • Right to an adequate standard of living, including adequate food, clothing and housing • Right to health • Right to education	• Right of peoples to: ○ Self-determination ○ Development ○ Free use of their wealth and natural resources ○ Peace ○ A healthy environment • Other collective rights: ○ Rights of national, ethnic, religious and linguistic minorities ○ Rights of indigenous peoples

	and have access to public office • Right to equality before the law and non-discrimination		

Obligations of signatory states

The obligations of the states-signatories of the documents within the IBHR are also 3 in number. These are the duties to *respect*, to *protect*, and to *fulfill* (Nowak et al., 2005; p. 11).

The obligation to *respect* generally means for the states to refrain from interference. That is generally taken as "prohibition of certain acts by Governments that may undermine the enjoyment of rights" (Nowak et al., 2005; p. 11).

The obligation to *protect* is understood as a requirement on behalf of states to protect individuals against abuses of their prescribed rights by non-state actors.

And finally, the obligation to *fulfill* means that states are "required to take positive action to ensure that human rights can be exercised" (Nowak et al., 2005; p. 12).

6.1.1.2 The European Convention on Human Rights

The European Convention on Human Rights (ECHR) alongside its Protocols is, arguably, the second most important document regarding human rights in the world. Its applicability covers almost the whole of the European continent (28 member states), with more than 510 million people (Statistics Explained, 2019), and including some of the most developed countries in the world.

The ECHR prescribes the following rights, freedoms, dignities, restrictions, limitations, and prohibitions:

- Right to life
- Prohibition of torture
- Prohibition of slavery and forced labor
- Right to liberty and security
- Right to a fair trial
- No punishment without law
- Right to respect for private and family life
- Freedom of though, conscience and religion
- Freedom of expression
- Freedom of assembly and association
- Right to marry
- Right to an effective remedy
- Prohibition of discrimination
- Derogation in time of emergency
- Restriction on political activity of aliens
- Prohibition of abuse of rights
- Limitation on use of restrictions on rights

Additionally, the ECHR also gives the legal padding for the establishment and functioning of the European Court of Human Rights. Interestingly, in Article 53, it explicitly excludes limiting or derogating from any human rights and fundamental freedoms which are binding and in force in any contracting party of the ECHR. All the existing obligations of a contracting party under, for example, arising from the IBHR documents continue to apply.

The Protocols to the ECHR deal with some additional prescriptions and protections:

- The Protocol signed in Paris in 1952 prescribes additional rights and protections: protection of property, right to education, and the right to free elections.
- The Protocol No. 4 signed in Strasbourg in 1963 adds the following: prohibition of imprisonment for debt, freedom of movement, prohibition of expulsion of nationals, and the prohibition of collective expulsion of aliens.

- The Protocol No. 6 signed in Strasbourg in 1983 deals with the abolition of the death penalty, and derogations in time of war.
- The Protocol No. 7 signed in Strasbourg in 1984 deals with the expulsions of aliens, the right of appeal in criminal matters, compensation for wrongful conviction, the right not to be tried or punished twice, and with equality between spouses in a marriage.
- The Protocol No. 12 signed in Rome in 2000 contains a general prohibition of discrimination.
- The Protocol No. 13 signed in Vilnius in 2002 deals with the abolition of the death penalty, and explicitly rejecting derogations in this sense.

Chapter III. Towards Ethics of Systems (the Metaethics)

1 Introduction

Why *metaethics*?

The label of *metaethics*[12] for this chapter was chosen because of the particular purpose of the material and the findings coming from it for the thesis. Therefore, it doesn't necessarily fit the 'traditional' definition of metaethics as commonly used in the literature on ethics. Instead, it fits the general definition as an "… attempt to understand the metaphysical, epistemological, semantic, and psychological, presuppositions and commitments of moral thought, talk, and practice" (Sayre-McCord, 2014). A similar approach is taken by DeLapp, by stating: "Whereas the fields of applied ethics and normative theory focus on *what is moral*, metaethics focuses on *what morality itself is*" [emphasis original] (DeLapp, 2019).

The approach is aimed at providing the needed metaethical basis for the development of an ethical framework for AI ethics, without utilizing circular or *ad hoc* argumentation. Necessarily, and similarly to Gödel's incompleteness theorems and especially Tarski's undefinability theorem, such basis needs to come from auxiliary, external source(s) that provide justification for the axiomatic ethical system comprised by the Ethics of Systems Framework.

This is where *systems theory*, and *ethics* and *philosophy of information* come. They both can provide the necessary background and justification for the building blocks on which AI ethics and the Framework itself can be studied, designed, analyzed, expanded, and utilized. Since in this work they represent the basis of its *ethics*, the name of *metaethics* was chosen as their category (and as natural part of the label of this chapter). The rationale is that both systems theory, and ethics and philosophy of information can provide an account for ethics in general, and from there, extend to AI ethics.

This approach in AI ethics is underrepresented, if at all utilized. Therefore, it is my purpose with this work to contribute towards the discovery of the metaethical foundations of AI ethics.

2 Ethics of Information

As I already mentioned before, *ethics of information* is a field which can provide significant contribution to the metaethics relevant for my work. One of the most developed takes in *ethics of information* currently is Floridi's (Floridi, 2013). This is the reason why I will mostly focus on this version (and related works, such as commentary, additional contributions, and criticism) in this work.

Before diving into it, however, we will need at least a cursory overview of its philosophical foundations—in *philosophy of information*.

2.1 Philosophy of information

Philosophy of information is a field of study focused on information, its nature, properties, and its place in the world. Floridi's own definition here is of great use, stating that:

> "**PI** The philosophy of information (PI) is the philosophical field concerned with (a) the critical investigation of the conceptual nature and basic principles of information, including its dynamics, utilization, and sciences; and (b) the elaboration and application of information-theoretic and computational methodologies to philosophical problems". (Floridi, 2011; p. 14).

[12] *Meta-* in ancient Greek: 'beyond' or 'after'.

Within the field there are several key concepts and tools (such as information and method of levels of abstraction), which I will explore here and which will be used further down this work.

2.1.1 Information

But, what is information? Surely any study that includes the concept of information has to provide some sort of a definition (and hopefully what information is *not*). Otherwise, there would be no way to make the concept useful.

Unfortunately (or, fortunately?), there is no lack of definitions coming from different fields, and even within the fields themselves. What information is for quantum mechanics, philosophy of information, systems theory, mathematics, and linguistics is not necessarily an identical notion—and at times differs wildly.

For example, Harshman (2016) provides a useful overview of how the notion of information shifted through developments in physics. Originally, information meant data that describes the current state of (part of) the universe, and knowledge of forces that propel the present into the future Harshman (2016; p. 8). This view presumed that total knowledge of the universe is achievable, or at least, a valid theoretical concept. However, statistical mechanics augmented the notion of information by introducing the notions of *macro-* and *microstates*. While a system can be well-modeled in simple terms in its macrostate (i.e. quantity of a gas in a container can be described by volume, pressure and temperature)—modeling its microstates (i.e. particular molecules) would require parameters in the order of 10^{24} to accurately describe (Harshman, 2016; p. 9). This also carries very important implications for the notions of *structure as constraint* (see 3.3.1.1 Structure is (causal) constraint below), the method of levels of abstraction as heuristic (see 2.1.6.1 Is the LoA method essentially a heuristic?), and bias in ethics (also see Chapter V. Discussion section 2.3.1 Substantial (ethical) implications).

Then, as the methods of analog and digital remote communication were developed, so the notion of information changed to a one based on *difference* (Harshman, 2016; p. 10). For example, the simplest unit of (digital or digitized) information was ‘found’ to be the bit, which can contain only absolutely *different* two states: 1 and 0. If a bit is 1 at a particular moment, it cannot (or at least it should not) be 0 and vice versa. Even in analog communication, transfer of difference is what creates the transfer of information, regardless if done in discrete steps. This is very close to Shannon’s notion of information in mathematical sense, which co-developed during this period. Additionally, the theory of relativity put an upper theoretical limit on the transfer of information—the speed of light.

Then, quantum mechanics ‘remixed’ the whole concept (Harshman, 2016; p. 11), by introducing a barrier to determinism (e.g. Heisenberg’s uncertainty principle), the apparent disruption of information by observation (e.g. the famous double-slit experiment and Schrödinger’s cat), the notion of informational and systemic coherence, entropy, (again) the knowledge from integration of the whole (macrostate) while information about parts (microstates) is missing, and the apparent faster-than-light transfer of information in quantum entanglement. These also bring important implications for the theory of integrated information (see 2.1.3 Integrated information), systemic structure and Being (see 3.3 Systemic Being), and other that I will discuss below in this chapter.

I included Harshman’s input here because it can provide us with useful overview of the conceptual variations surrounding information. But there are other notions of information that are not physical in nature. For example, Searle insists on claiming that information is observer- or consciousness-relative (Searle, Tononi & Koch, 2013). He is probably focused on the notion of *semantic information* (see below), which implies understanding—for which an observer or consciousness is needed. Data can only be meaningful to some*thing* that can extract meaning out of it.

2.1.1.1 SEMANTIC INFORMATION

Which brings the discussion to *semantic information*, or, information as a vehicle to convey meaning. It is the connection between data and meaning extracted out of it. Since I will be working mostly with Floridi's version of *ethics of information* here, it would be best to use the General Definition of Information (GDI) included in his work (Floridi, 2011; p. 84). Therefore,

> "GDIσ (an infon) is an instance of semantic information if and only if:
>
> GDI.1 σ consists of *n data* (d), for n ≥ 1;
> GDI.2 the data are *well-formed* (wfd);
> GDI.3 the wfd are *meaningful* (mwfd = δ)".

He goes on to break down the meaning of the elements of the GDI:

> "According to GDI.1, semantic information comprises data. We shall see that things can soon become more complicated. In GDI.2, 'well-formed' means that the data are clustered together following the rules that govern the chosen system, code, or language being analysed. Syntax here must be understood broadly (not just linguistically), as what determines the form, construction, composition, or structuring of something. Engineers, film directors, painters, chess, and gardeners speak of syntax in this broad sense. As for GDI.3, this is where semantics finally occurs. 'Meaningful' means that the data must comply with the meanings of the chosen system, code, or language in question. In this case too, let us not forget that semantic information is not necessarily linguistic. For example, in a map, the illustrations are such as to be visually meaningful to the reader".

However, according to Floridi this is not enough. This is enough to describe semantic *content*, but for it to become semantic *information* it needs a final component—truth-constitution (veridicity). With this in mind, basically, information is "well-formed, meaningful, and veridical data" (Floridi, 2016b).

Taken formally,

> "DEF   *p* qualifies as factual semantic information if and only if *p* is (constituted by) *well-formed*, *meaningful* and *veridical* data" (Floridi, 2016b).

This is the definition I will be using the most in the work below.

Why is the concept of information important for this thesis?

As mentioned before, ethics is in large part comprised of the gathering, utilization and manipulation of data and information. Entities, when pursuing their goals in the world, rely on data and extracted information out of it to guide themselves and make the best of available resources. This information can be about the external world (including other entities), but it can also be about their internal world e.g. how much resources they have, how they feel, how estimated risk fares against potential gains for an (in)action, are they tired, disappointed, angry, and similar. Additionally: moral and legal rules, moral status, duty to do *x*, and not do *y*, 'maximize happiness', 'avoid suffering', an ally, a friend, an enemy … these are all mental concepts (complexes), based in data and information.

Essentially, an entity looking at the world and making sense of it is exactly an informational process. This applies equally to complex and simple entities, to both 'natural' and 'artificial', to collective and unitary. Generally speaking, the more complex an entity is usually translates into increased capacity to handle data, extract information and use it.

2.1.2 The infosphere

The ‘infosphere’ is Floridi’s term to conceptualize the end result of the transformational process caused by Information and Computer Technologies (ICTs) over the world. This end result is a blend between what is physical and what is digital, a sort of re-ontologized world (by ICTs)[13] in which there is no difference between the two. One can recognize this re-ontologized world by references to the Internet of Things (IoT), Internet of Everything (IoE), ambient intelligence, and similar notions.

In the infosphere all entities that comprise it would necessarily scale down to a view focused on the informational level of abstraction (see 2.1.6 Method of Levels of Abstraction below)—which shows all entities as *informational entities* (see below).

Basically, the world is seen as (or becoming) the infosphere, comprised by all entities that exist (which are informational entities) and their properties, interactions, processes, and mutual relations (Floridi, 2013; p. 6).

2.1.2.1 Informational entities

As mentioned above, the infosphere is comprised of all entities. Since it is the infosphere, entities are (seen as) *informational entities*. They are also sharing the same informational nature (Taddeo, 2016; p. 368). An informational entity is a “consistent packet of information, … , an item that contains no contradiction in itself and can be named or denoted in an informational process” (Floridi, 2013; p. 65).

This last point on contradiction is important. According to Floridi, an entity is a delimited object—part of the infosphere, which is defined by having no contradiction in itself. To be an entity, a packet of information must not positively involve a contradiction, since it will end up being a further source of contradiction, and will become a case of total negation—an “informational black hole”. Thus, there “… are no information processes that fruitfully involve contradictions (obviously this is not to say that there are no contradictory information processes), that an information process can involve anything which is in itself logically possible, and that IE treats every logically possible entity as an informational entity” (Floridi, 2013; p. 65).

Spoken in the language of logic, if A is an informational packet and contains:

$A \land \neg A$; or simply: $\neg A$ (with A already implied);

then A is contradictory and cannot be an informational entity.

However, this strikes me as a too restrictive interpretation of Being and entities. Defining entities in this manner might stem from the logico-informational point of view of hard delimitation that Floridi follows, which is typical of Western philosophical thought; but which is too restrictive and ‘empty’ according to some Eastern philosophical traditions e.g. Buddhism and Daoism. We will see further down in the systems theory section that entities (systems) can be defined more widely, in a positive (*is-ness*) and negative (*is-not-ness*, difference) manner, and this can be more accommodating even to entities that are contradictory in themselves—but are still Beings nonetheless.

For now, though, the notion of consistency is a point that is, and will become very important for the Ethics of Systems approach I formulated in this Chapter (see 4 Towards Ethics of Systems below).

[13] This has the odor of strongly biased anthropocentrism, when and if we believe that only the world we humans and our extensions (e.g. ICT technologies) are able to perceive is the whole world—which is far from true.

Being as pattern

Bynum also has something to add to the discussion, by commenting on the revelations of MIT mathematician Norbert Wiener. Wiener had a revelation that entities and processes in the universe can be viewed as "… patterns of information (data structures) encoded/embodied within an ever-changing flux of matter-energy. Every physical object or process is part of a creative coming-to-be and a destructive fading-away, as current information patterns – data structures – erode and new ones emerge. This "Wienerian" view of the nature of the universe makes every physical entity a combination of matter-energy and physical information" (Bynum, 2016; p. 205).

He then goes on to say that

> "Even living things, according to Wiener, are informational objects. They store and process physical information in their genes and use that information to create the building blocks of life, such as amino acids, proteins and genes. Indeed, they even use stored information to create new living things; namely, their own offspring. Animals' nervous systems store and process physical information, thereby making their activities, perceptions, and emotions possible. And, like every other physical entity in Wiener's universe, *even human beings can be viewed as informational entities*. Thus, humans are essentially *patterns* of information that persist through an ongoing exchange of matter-energy" [emphasis original] (Bynum, 2016; p. 205).

I discuss the view of Being as pattern in more depth below in 3.3.2 Being as pattern.

2.1.3 Integrated information

Integrated Information Theory (IIT) is a recent (Tononi, 2004) addition on the theory of consciousness and mind proposed by the neuroscientist and psychiatrist Giulio Tononi, and further developed and advocated for by his collaborators (most notably Christof Koch) and others (Barrett, 2014) (Horgan, 2015) (Oizumi et al., 2014) (Mørch, 2018). IIT is also subject of critique, for example by Cerullo (2015) and maybe most notoriously, by John Searle who was prone to comment that IIT "… does not seem to be a serious scientific proposal" (Searle et al., 2013). Currently IIT is at version 3.0.

IIT is a novel and systemic approach at explaining consciousness that starts in inverse direction from the typical one in neuroscience. It starts by identifying the essential, self-evident phenomenal properties of consciousness (which are dubbed *axioms*), and then attempts to identify the correlates of these axioms—the physical properties of systems that can account for them (dubbed *postulates*) (Tononi, 2015).

By talking about systems[14], IIT is one of the two major bridges between systems science and philosophy of information. It is also a particular kind of panpsychism, because it posits that an integrated system holds, or **is**, a particular kind of experience *in itself*—an experience of what it is to 'be that system' (Koch & Tononi, 2014). This makes it particularly fitting for the work here.

2.1.3.1 IIT's components

IIT's *axioms* are ***existence*** (consciousness exists); ***composition*** (consciousness is compositional, and each experience consists of different components); ***information*** (consciousness is informative, and each experience is defined by how it differs from any and all other experiences; see also 2.1.5.1 Reasoning as the perception and manipulation of differences below); ***integration*** (consciousness is integrated, each experience is strongly irreducible to non-independent components); and ***exclusion*** (consciousness is exclusive; each experience excludes all others at a given time) (Oizumi et al., 2014; p. 2).

[14] Although IIT defines a system as a "set of mechanisms" (Oizumi et al., 2014; p. 3); which is different from, but compatible with, the definitions used in systems theory (see 3.1.1 A system below).

IIT's *postulates* are ***existence*** (mechanisms exist, and systems are sets of mechanisms); ***composition*** (mechanisms can be organized in higher-order ones); ***information*** (a mechanism can only contribute to consciousness if it represents a "difference that makes a difference" within a system); ***exclusion*** (a mechanism can contribute to consciousness only a single, integrated (irreducible) cause-effect repertoire).

It would be good to add here that I don't support the strong *exclusion* postulate that is posited by IIT, and I advocate for a softer exclusion that increases IIT's explanatory power over some panpsychistic phenomena (see the next section on panpsychism for a more detailed discussion). The exclusion postulate also is subject of other modifications and commentary, albeit for different reasons, in Mørch (2018).

IIT also offers a way to measure the level of integrated information of a system (at least in principle), and thus discern a particular experiential state (consciousness) from another. But difference can also be small, or even zero, which would mean that two systems that contain the same integrated information have a similar or even equivalent experience, simultaneously or at different points of time. The maximum value of integration is noted as φ^{Max}. A local maximum of integrated information (over elements, time and space) is noted as Φ^{Max}.

IIT specifies that "an experience is thus an intrinsic property of a complex of mechanisms in a state" (Oizumi et al., 2014; p. 3). For this purpose, it deals with two concepts: maximally irreducible cause-effect repertoire (MICE) and maximally irreducible conceptual structure (MICS). Experience (consciousness, phenomenology) is equal to the MICS of a system in a qualitative sense, and it can be extrinsically measured with the Φ^{Max} measure. On the other hand MICE specifies what is a concept, and what it contributes to the quality of experience. MICE can be extrinsically measured with the φ^{Max} measure.

Furthermore, they add that "In other words, the maximally irreducible conceptual structure specified by a complex exists intrinsically (from its own intrinsic perspective), without the need for an external observer" (Oizumi et al., 2014; p. 3). This represents a kind of a response to Searle's comment that information is "only information relative to a consciousness. Either the information is carried by a conscious experience of some agent (my thought that Obama is president, for example) or in a nonconscious system the information is observer-relative—a conscious agent attributes information to some nonconscious system (as I attribute information to my computer, for example)" (Searle et al., 2013).

Although Searle's comment is correct (information does depend on an entity giving meaning essentially by providing or creating relations—i.e. systemhood—to a set of data), according to IIT the experience (the content of consciousness) of a system *is the information itself* because the entity gives meaning to its experience *res ipsa loquitur*; and this is not dependent on any external observer. The internal observer is, or is 'devising', the information. Otherwise, we would have to accept that an internal observer is either impossible to exist (there is no such thing as subjectivity), or that the process of creating or extracting meaning out of data is a 'magical' one not connected to any material processes and states in its 'brain' or elsewhere.

Table 3: Key concepts and measures of IIT (adopted from Oizumi, Albantakis and Tononi, 2014)

Mechanism	System of mechanisms
Information Only mechanisms that specify differences that make a difference within a system count	
Cause-effect information (cei): How a mechanism in a state specifies the probability of past and future states of a set of elements (cause-effect repertoires)	**Conceptual information (CI)**: How a set of mechanisms specifies the probability of past and future states of the set (conceptual structure)
Integration Only information that is irreducible to independent components counts	
Integrated information (φ, ‘‘small phi’’): How irreducible the cause-effect repertoire specified by a mechanism is compared to its minimum information partition (MIP)	**Integrated conceptual information** (Φ, ‘‘big phi’’): How irreducible the conceptual structure specified by a set of mechanism is compared to its minimum information partition (MIP)
Exclusion Only maxima of integrated information count (over elements, space, time)	
Concept (φMax): A mechanism that specifies a maximally irreducible cause-effect repertoire (MICE or quale “sensu stricto”)	**Complex** (ΦMax): A set of elements whose mechanisms specify a maximally irreducible conceptual structure (MICS or quale “sensu lato”)

Implications and importance of IIT for this work

Being, structure, and identity

The first, very important implication of IIT for this thesis is regarding structure, Being and identity.

IIT stipulates that there is “… an identity between phenomenological properties of experience and informational/causal properties of physical systems” (Oizumi et al., 2014; p. 3). What this means is that the subjective qualia of experience (consciousness) of a particular system is not only directly correlated with, but is its physical structure and the relations between its components. An entity’s Being (identity, or at least its internality in a weaker sense) is the direct result of its systemic structure—and now comes the moral perspective—the breakdown of systemic structure means breakdown of an entity’s Being (identity). Since, as we will see in 4.3.2 The moral imperatives below, one of the two moral imperatives is the **Conservation of Personal Continuum** (**CPC**; which can be roughly equated with Being/identity of an entity), it follows that how much CPC is respected depends on how much a system’s integrated information is conserved.

Cognition and reasoning

The theory of integrated information also carries on another, interesting and very important implication regarding cognition and reasoning.

If we take that a system has integrated information in itself, which is based on the *particular* configuration of its components and their relations (see 3 Systems theory below), if there is a change in that particular configuration, at the same time, there is a change in the integrated information. A system that changed in its structure, also changed in the integrated information it holds inside its (subjective) state.

Now, let’s imagine a system that has a special ‘information dealing’ subcomponent. This can be just a part of the system which is dealing with extracting information out of available data (similarly to how some parts of the brain are dealing with particular tasks especially regarding the senses i.e. visual, auditory, olfactory, tactile, temperature, pain and other data). In this case, the major systemic complex is not tasked with extracting information, but may be spurned into paying attention if the information dealing subcomponent serves information that is *out of the ordinary* (i.e. which represents a significant change from a previous state).

The important thing to note here is *change*. Change in a system's structure creates change in its integrated information, and by this, in its mental state. While a system can (theoretically) remain indefinitely into an unchanged structural state, typically it doesn't. But change can be good, as it can enable a new kind of experience—the so called *access consciousness*—which in turn enables reasoning as a more complex type of experiencing (see 2.1.5 Cognition and reasoning below).

Panpsychism

And the third implication of IIT is regarding panpsychism.

The aforementioned makes a natural jump to panpsychism in the next subsection. The integrated information in a system as a baseline, 'pure' level of consciousness without content (Oizumi et al., 2014; p. 17) constitutes a particular experience that exists all while the system continues to be integrated in such configuration. If all systems contain integrated information and thus experience within themselves, then all systems have a mind i.e. that experience *is* the mind.

2.1.4 Panpsychism

Panpsychism is the philosophical position that claims that "… the components of the world have some inherent experiential or mind-like qualities" Skrbina (2009a; introduction).

Panpsychism often comes from the quest to find the origin of mind. Typically, philosophers that discover or invent panpsychism do so after trying to determine how does mind get formed from the apparent physicality of the Universe. Strict physicalists typically hold that the nature of reality and the substance of the universe are strictly physical. But that has the annoying implication that mind cannot be possible to exist—and we know from our personal *cogito ergo sum* experience that it does (at least our own mind). Otherwise they are left to explain the magical appearance of mind seemingly out of nothing (Goff, 2009). This appearance out of nothing is what proponents of *strong emergence* support (see 3.1.3 Emergence below). However, we seem to be meeting only *weak emergence* in nature (Spät, 2009; p. 164).

Hence, panpsychists have concluded that a mind-like property must be intrinsic to the building blocks of the Universe, regardless if they are physical in nature. Or, as Goff would say: "It seems like we don't need to explain where consciousness came from if it was there all along" (Goff, 2009; p. 130). This mind-like property can combine, fuse, or defer its mind to higher-level forms, such as minds created by biological brains. What's more, according to some authors like Freya Matthews, the panpsychist paradigm might actually be the one coming before the physicalist / materialist one, and be responsible for the actual creation of the physical universe (Mathews, 2011).

We can also immediately notice that there is no intrinsic, *a priori* barrier to the possibility of a mind forming out of any *other* material type in the Universe—including 'artificial' ones such as silicon and thus derived logical gates in computers. This is the position implicitly supported by IIT, which I covered in the previous section. This is also why panpsychism is important for this thesis (see below).

If the nature of reality (including its physical aspect) is panpsychist (i.e. there is a mind-like property in everything), an integration of a set of things into a system would automatically translate to integration of (parts or contributions of) the separate minds of the components into a new, unified mind (consciousness). Thus, Diderot:

> Diderot went further, tackling the combination problem and the unity of mind. On his view, if particles of matter are sensitive and intelligent, then simply by virtue of communication and contact they can form an integrated being. He made an analogy with a swarm of bees: "This cluster is a being, an individual, an animal of sorts." (…). It is a unitary being because of the extremely tight interaction between parts, which pass from being merely "contiguous" into being truly "continuous." The human body is similar to the swarm of bees; the body is a collection of organs, which "are just separate animals held together by the law of continuity in a general sympathy, unity, and identity." It is the "continual action and reaction" between parts that creates the unity; "contact, in itself, is enough" (…) (Skrbina, 2009b; p. 14).

This, however, does not mean that necessarily the *whole* mind of the systemic components is integrated. They retain some personal, subjective mind with a personal, delimited point of view; but also contribute part of their mind to form the mind of the unified whole (see 2.1.4.1 Soft exclusion below; also, Mathews (2011; p. 5). Or, alternatively, the mind of the whole emerges out of the minds of its components, and is not directly connected to, or the *direct result* of, them.

Hence, separate neurons of the brain might not be aware that there are neuronal clusters (brain modules) to which they belong. They individually function as is necessitated by their personal systemic structure. This also means that they are individually conscious of a limited aspect (in volume, amount, substance) of the information flows that take place throughout the whole brain. Separately, this way of functioning performs and contributes to a separate, higher-level function of the cluster they belong to.

Furthermore, the clusters might not be aware of the existence of a unified whole brain to which they belong and for which they perform a certain function (for example tracking of objects through the eyes by the visual cortex, integration of memory in the hippocampus, tracking of rhythm and pitch in the auditory cortex, and similar). Additionally, that brain might not be aware of the higher-level mind of the whole body, and the body might not be aware of the higher-level mind of a group of people, and then a company, a nation, an ecosystem, a continent, the planet … ending with the universe itself (similarly, see Smith (2017; p. 25)).

2.1.4.1 **Soft exclusion**

What was last said would seem to go directly against the (hard) *exclusion* postulate that Oizumi, Albantakis and Tononi defined in their IIT (Oizumi et al., 2014). I already mentioned before (in 2.1.3.1 IIT's components above) that I am in favor of a soft(er) exclusion postulate, rather than the hard one endorsed by IIT's authors—as a solution to the so-called 'combination problem' (Goff, 2009; p. 130).

According to IIT, when a set of elements of a system (or the whole system itself) are integrated into a complex—a maximally irreducible conceptual structure—only that integrated set *as a whole* contributes to consciousness. The components that build it simply cease to contain consciousness, and in a sort of 'wizardly' way transfer their, let's call it consciousness potential, to the integrated whole. In simpler terms, if consciousness of an integrated whole arises, it *excludes* the possibility of consciousness in its components. This is the stance of the panpsychist flavor known as *fusionism* espoused by philosophers such as William Seager and Hedda Hassel Mørch (Goff, Seager & Allen-Hermanson, 2017).

But this need not be. I will here try to show that exclusion indeed does occur, but in a hierarchical and communicative fashion. What do I mean by this?

Let's imagine a bundle of components of a complex system, such as a crowd of protesters. At times the crowd is simply a non- or disintegrated set of individuals that are all strictly aware only of their own minds, and choose to participate in a purely individualistic fashion.

However, social psychologists have noticed the so-called ‘herd mentality’ phenomenon, or ‘herding’ in humans (Raafat, Chater & Frith, 2009). When a set of people seem to change their behavior in a more synchronized fashion with the entire group, especially while lacking centralized command-and-control, they are engaging in herding behavior. In the words of the authors, “herding can be broadly defined as the alignment of thoughts or behaviours of individuals in a group (herd) through local interactions rather than centralized coordination. In other words, the apparent central coordination of the herd is an emergent property of local interactions” (Raafat et al., 2009).

Herding behavior is particularly suited because it will convey my argument in a plastic manner. When a crowd becomes a herd the behaviors of its components synchronize and start to act as if ‘possessed’ from an outer source of control. No individual there becomes aware of the mind of the *whole* herd, because no individual has access to all the information (mental content) that are available to all other individuals. They only have access to direct experiential information, and information which is communicated and thus indirect. Thus, they are *excluded* from all other information (which is why partial exclusion takes place).

However, the herd itself has access to information on the level of the herd, even though it doesn’t have access to the totality of consciousness of the individual components. What is going on here? I am willing to argue that the herd emerges out of the crowd by integration of its components (mostly the people in the crowd), and thus gains its own mind. This emergent mind guides the ‘possessed’ behavior of the individuals inside it (which can include all or part of all the individuals in the set). At that particular moment the herd is an integrated system, and it fills the individual minds of its components with the collective, ‘herd’ content; but of course, split as per the position and communicative relatedness of the particular individual inside.

This, however, does not imply that the individual suddenly loses its integration (which would be stipulated by the hard *exclusion* postulate), loses its consciousness, cedes the totality of its mental powers to the herd, and thus becomes a zombie—even temporarily. The individual retains its local integration, but the *content* inside its mind are partially or fully replaced with the one served by the herd through its channels of communication with that particular individual. Since the content is integrated and spread across multiple carriers, it makes no sense when looked upon partially i.e. if we look upon the content present in a herd member, it will make little to no sense *on its own*. Similarly, Mathews (2011; p. 5): “Selves then enjoy a real though relative individuality even though they exist in the context of an undivided whole. Since they proactively seek from their environment the resources they need to actualize and maintain their structure while at the same time resisting causal inroads into their integrity, they count, ontologically, as individuals, even though they are not separate substances, but disturbances within a global substance”.

Hence, at the same time both the herd *and* its individuals can be integrated. The *soft exclusion* here plays a role to exclude access of the individual to most of the information (mental experiential content) that exists in the herd, and vice versa. Interestingly enough, IIT seems to be open to soft(er) exclusion, by mentioning the possibility of elements of a system being ‘para-conscious’ besides contributing to the higher-level consciousness (Oizumi et al., 2014; p. 17).

Other examples can also convey the validity of soft exclusion. For example, a water pump does not cease to be that particular and integrated water pump when it is installed in a heating system in a building. The pump continues to be integrated and to function in a particular manner, but at the same time contributes to create something more—the heating system.

When we connect a computer with others to create a networked computer system, the computer does not loose local integration nor the ability to process inputs and give out outputs as before according to its internal structure. However, the content of the inputs and outputs will probably become different, and the computer will *additionally* integrate itself into the larger system while retaining local integration.

Similarly with nations, collectives, organs of both biological bodies and states, companies, planets, and all components of the Universe that are or become locally integrated. They can all potentially become further integrated into a larger whole, while retaining local integration. Thus, they will contribute to a larger mind while also retaining their own 'local' mind, which might be partially or fully filled with content from the whole to which they belong to.

This clearly has panpsychist implications.

Why is panpsychism important for this thesis?

Depending on what we take to be important for an ethical theory, panpsychism can have clear moral implications. The most straightforward example would be with what Floridi calls the *uniformity of Being Floridi (2013; p. 65)* which is a position that I also support here (see 4.3 The moral entity). If Being is taken as morally valuable, we can derive the conclusion that the its uniformity is also valuable.

The intrinsic value of Being

Being can be taken as having both external and internal properties. An external property is what is Being when described by objective, external factors, such as structure, structural resilience, space, time, cause-effect repertoire, components, relations, and similar. However, Being also has internal properties (i.e. aspect). According to panpsychism and IIT there is something like being a particular entity (system). Being is an integrated constellation of matter-energy that has a certain subjectivity, a 'mind'. For example, experiencing the color red on a piece of paper is an internal property of me as a human being which also has external properties (e.g. a body). In this sense, if we value (the uniformity of) Being then we value (the uniformity of) this subjective existence.

Conservation of Personal Continuum

At the same time we ought to be aware that *total* conservation of Being (of both its external and internal properties) is not only improbable, but also impossible. The apparent inherent stochasticity of an entity's environment consistently instills changes to the external properties of that entity, and with that, to its internal properties. If a moral theory aims to offer realistic guidance on what we ought to value and do, it has to take into consideration this, also.

The most we can realistically aim to conserve, then, would be the *continuity* of Being through time and space. It is hopeless to keep an entity absolutely intact through the times, but we can instead aim to preserve the continuum from one to another point in time and space. This lends support for one of the two moral imperatives, named **Conservation of Personal Continuum** (see 4.3.2.3 Conservation of Personal Continuum in this chapter).

For example, it would be foolish of me to expect to remain the same person that I have been 10 years ago, both morally and in general. But it would not be foolish of me to aim at keeping my continuity through space and time i.e. to avoid death, both self- and other-inflicted. The first 'hope' cannot be integrated in a realistic moral system, while the second aim is typically one of the most supported moral and legal principles (i.e. dignity of life and human rights in general).

The jump from ‘is’ to ‘ought’

Spät similarly claims that panpsychism can enable making a jump from ‘is’ (physical reality) to ‘ought’ (experiential reality). His claim is in contrast to authors such as Wittgenstein, who claim everything that happens in the universe just happens, and there is no such thing as value. Contrary to this, Spät argues, a panpsychist universe entails *intrinsic value* in things (systems and entities). All things have a possibility to suffer, and potentially a ‘desire’ to avoid suffering. Out of this we can derive the, albeit weak, conclusion that avoiding suffering (‘negative’ experiential state) is a desired direction in which we *ought* to try and move towards[15]. This, as Spät says himself, “the ‘is’ entails the ‘ought’” (Spät, 2009; p. 170).

This *intrinsic value* seems to guide at least some human moral intuitions i.e. natural and human rights theories, environmentalism and deep ecology, ethics of care and similar patient-oriented ethics, as well as (some flavors of) deontology.

Dignity of life

Therefore, it lends credence to the concept of *dignity of life*. As Weber and Varela say, “in observing other creatures struggling to continue their existence – starting from simple bacteria that actively swim away from a chemical repellent – we can, by our own evidence, understand teleology as the governing force of the realm of the living” (Weber and Varela, 2002; in Spät (2009; p. 171)).

Moral dignity for informational entities (which can, for most purposes, be equated here with existence in a panpsychist sense) is also a principle followed by Floridi in his *Ethics of Information*: “... IE holds that every informational entity, insofar as it is an expression of Being, has a dignity constituted by its mode of existence and essence, defined here as the collection of all the elementary proprieties that constitute it for what it is. This dignity *prima facie* deserves to be respected and hence may place moral claims on any interacting agent. It ought to contribute towards constraining and guiding her ethical decisions and behaviour, even if only initially and in an overridable way” (Floridi, 2013; p. 69).

Soft exclusion, and individual and collective rights and dignities

The soft exclusion principle that I discussed above can have some additional moral implications. If a component of a system can remain both locally integrated and also in the bigger system, that would mean that moral respect ought to be given both to the local and higher-level integrations (Beings). That would also give rise to the need to do a balancing act between these two or more (levels / types of) dignities, thus rendering morality as a balancing act itself (see below).

For example, a person belonging to a group, nation, or a culture, can both be locally, personally integrated (by being an individual); but also he can be integrated into the higher-level system. Which integration is ‘more’ important? Can the higher-level system simply use the individual for its purposes, disregarding its dignity (like some collectivist philosophies advocate for e.g. the tried and failed socialist ‘utopias’ of the 20th century)? Or can the individual disregard the dignity of the system it belongs to for his own, selfish purposes (like some individuality-based philosophies advocate for i.e. laissez-faire capitalism and libertarianism)?

Humans seem to have reached a point where we discovered that *both* these integrations must be considered equally important if we want to put suffering, war, conflict, and instability to a minimum. We have achieved this by creating the framework of the United Nations and the universal human rights and dignities, who not only prescribe individual, but also collective rights; and also democratic governance and respect for the rule of law (the rules of the higher-level system to which individuals belong, such as a state).

[15] This seems to support the idea I presented before in 3.1.3 Good as the absence of bad, that moral Good might not be a positive state of matters; but, on the contrary, the absence of moral Bad/Evil e.g. the avoidance of suffering.

Systemic goals and the Achievement of Personal Goals

This ‘teleology’ that Weber and Varela talk about can be understood generally as the propensity of systems (entities) to aim at achieving goals (goal = telos; gr. τέλος); and specifically, when talking about biological nature, as the propensity of living beings to aim at continuing their existence, at least until they achieve their (other) imperative goals. And, continues Spät, where “there are ends, goals, and purposive behavior, there are values: The purposive ‘is’ implies an ‘ought’, i.e. the purposiveness implies *values*” (Spät, 2009; p. 172). This is for the simple obvious reason that having goals means that a system considers something as valuable to be pursued, protected, or conserved (and spend resources doing this); and the criteria that are used, implicitly or explicitly, to determine what is and is not valuable are called *values* (see 3.1 Value theory in Chapter II. Literature review above).

This goal-seeking behavior of systems also lends support to the second moral imperative that I included in this work, the **Achievement of Personal Goals** (see 4.3.2.2 Achievement of Personal Goals below). Thus, argues Spät, “the purpose to live and to survive is an intrinsic property of physical reality” (Spät, 2009; p. 172).

Morality as a balancing act

A final point that I will include from Spät’s article is a very important one. According to him, even though panpsychism can entail respect for the dignity of life, we can go overboard with it. For example, even though a cancerous cell is also a living being, we would be fully morally allowed to destroy it so that it doesn’t destroy us. Another example is that we need food (i.e. fruits, vegetables, and some amount of meat) to survive and thrive. In short, we “… need to do harm to e.g. the purposiveness of fruits and vegetables in order to ‘satisfy’ our own purposiveness, i.e. in order to survive” (Spät, 2009; p. 174).

This brings us to a moral dilemma: when are we allowed to harm another being (system, entity) for our purposes?

Some (proto)panpsychist philosophical positions (i.e. Mahayana Buddhism) have tried to develop the principle of avoidance of suffering to an extreme end, ending up with the conclusion that we are not allowed to purposely harm a living being at all. This translates also into being obliged to be a vegetarian.

However, settings the threshold for living beings just after plants and before animals strikes me as very arbitrary. True, typically plants do not act as fast or in as conscious manner as animals, but they are living beings nonetheless. Some plants are way more complicated and aware of their surroundings than many viruses, bacteria, and some simple kind of animals. What gives us moral permission to eat and destroy plants, but not animals? What about ecosystems, mountains, rocks, rivers, planets, or star systems? Are we allowed to use and/or destroy them because they (*seem*) less aware, conscious, and alive?

It seems that the direction we can take is one of satisficing (see 4.1.1.2 Consequentialism (teleological Ethics) in Chapter II. Literature review). Namely, we are allowed to “satisfy our purposiveness” by using other beings, but only up to the point of survival and mindful flourishing. We ought to balance our survival and flourishing with survival and flourishing of other Beings, and do our best to “protect the dignity of life as far as possible” (Spät, 2009; p. 175). He thus re-formulates a ‘categorical imperative for the dignity of life’ (reformulation originally offered by Jonas) in the following, obviously Kantian, form:

> Act so that the effects of your action are – as far as possible – compatible with the permanence and the dignity of life (Spät, 2009; p. 175).

In this sense we are clearly not allowed to harm another being just for the sake of it. Causing destruction and suffering without any morally righteous cause is simply morally impermissible, and many times deeply morally repulsive.

I will discuss another principle that Floridi uses in his work to determine whose (informational) entity's dignity has preference to another and in which situation further down below in 2.2 The ethics of information.

2.1.5 Cognition and reasoning

It is intuitive to include cognition and reasoning in a section about philosophy of information, but why do I include it at all in this work? As I mentioned before in Chapter II. Literature review, section 3. Moral reasoning, there can be no ethical theory that does not propose at least the basic tenets of how moral reasoning ought to function in moral scenarios.

Moral reasoning is a 'special' kind of reasoning that deals with moral calculus. We (humans) perform our moral reasoning on an everyday basis when we ponder: whether to steal that unguarded bag someone left; whether to follow the legal and moral rules of our communities and society; whether to do what is right instead of what is easy, manipulative, and plainly wrong. Whatever (we know that) produces morally-burdened effects by our decision-making necessarily employs moral reasoning.

AI entities will also produce morally-burdened effects. The more they are integrated with our lives, the more significant these effects will become. If we want to manage these we will need to implement moral decision-making in these systems.

This can be done in two ways:

a) Implicitly—the spectrum of decisions that produce morally-burdened effects is explicitly known by designers, implementers, employers, or users of the systems. Moral theories are used by the aforementioned to make the systems choose the 'right thing', even though the systems do not follow any specific moral theory nor do moral calculus themselves;

b) Explicitly—the AI systems themselves have a moral decision-making capacity implemented in them, making them capable of performing explicit moral calculus autonomously or semi-autonomously.

But what are cognition and reasoning?

Oxford's Lexico defines cognition both as a process, and as an end product of that process. Hence, cognition is "the mental action or process of acquiring knowledge and understanding through thought, experience, and the senses". It is also "a perception, sensation, idea, or intuition resulting from the process of cognition" (Lexico, 2019a).

Reasoning is also, on the other hand, a "mental action or process", but is focused on "thinking about something in a logical, sensible way" (Lexico, 2019b). So, reasoning follows implicit or explicit logical structure, in order to make the inferential jump from propositions (e.g. reasons) to a conclusion. Therefore, reasoning is a subset of cognition, even though they are often conflated.

Cognition is not always reasoning, since biological systems (such as humans) are obviously capable of perception and understanding of their environment even though they typically do not follow formal or explicit rules of reasoning (e.g. logic).

But reasoning is important for this work since it enables us to explicate the cognitive process that would be utilized to understand and manage moral scenarios, both by people and organizations that seek to manage them, and by AI entities that participate in them. Therefore, it is best to try and understand what reasoning is itself.

2.1.5.1 Reasoning as the perception and manipulation of differences

Reasoning is a cognitive process where implicit or explicit rules are followed with the aim of reaching a final product—conclusion. These rules serve to guide the process of making an inferential jump from premises to the conclusion.

In the most basic sense, even with a single premise, there has to be a *difference* between the premise and the conclusion; otherwise the process makes no sense. Moreover, if there are multiple premises there also has to be a difference between them so that it makes sense to speak about *multiple* or *different* premises.

For example, we can make the simplest logical calculus with a single premise, *a*, and the 4 basic logical connectives in CPL: $\wedge$, $\vee$, $\neg$, $\rightarrow$. If we simply state *a*, then we can redundantly say that *a* implies *a*[16] by itself,

$a \rightarrow a$.

Or, in other words, if *a* then *a*.

Even here there is a difference between the premise (*a*) and the conclusion ($\rightarrow a$), since it would not make sense to talk of a conclusion (implication) if there was no difference between the two, even for redundant calculus such as this.

However, *a* cannot exist in a vacuum. *a* by itself contains nothing meaningful. *a* is *a*, because everything else <u>*is not*</u> *a*. Thus here we can explicate 'everything else' as not *a*, or: $\neg a$. We can even take the meaning of 'everything else' and assign it to *b*. Therefore,

$b = \neg a$;

but also—if simply *b* then not *a:*

$b \rightarrow \neg a$.

Finally, if *b* then not *a*, then, if *a* then not *b*:

$(b \rightarrow \neg a) \rightarrow (a \rightarrow \neg b)$

Here it becomes more obvious that difference is in the foundations of reasoning[17], since without difference it makes no sense to talk about reasoning steps, operators, premises, and conclusion. These are evidently *different* concepts.

Even without following formal rules, a cognitive entity is not able to perform cognition if there is no difference between one or another state of matters (i.e. between food and not food, danger and not danger, light and dark, etc.). Similarly, Harshman comments that the switch is the simplest electrical device; the flow or absence of current describes a basic binary piece of information: a bit (Harshman, 2016; p. 10).

Difference seems to be the foundational property that enables cognition, and specifically, reasoning. Ashby (1999; p. 9) thus claims that, for example, in cybernetics the most important concept is that of difference. In

[16] This (self-reference) is what Heinz Von Foerster comically refers to as the forbidden land where the "devil's cloven-hoof in its purest form" resides (Von Foerster, 2003; p. 289).

[17] This also seems to be implied in Russell and Whitehead's introduction of negation (¬) as the first connective, and further commenting that negation is necessary to bind more strongly than all the other 3 connectives (Bezhanishvili & Fussner, 2013; p. 3 and 4).

CPL difference is most simply represented by negation: not, ¬. Negation also further enables disjunction ∨ (or) and conjunction ∧ (and).

It is then understandable that when dealing with data and information, difference would be one of their foundational properties. Likewise, Floridi also takes difference as the defining property of data and information:

> The fact is that a genuine, complete erasure of all data can be achieved only by the elimination of all possible differences. MacKay (1969) highlighted this important point when he wrote that 'information is a distinction that makes a difference'. He was followed by Bateson (1973), whose slogan is better known, although less accurate: 'In fact, what we mean by information—the elementary unit of information—is a difference which makes a difference.' Total data erasure means the erasure of MacKay's 'distinction' or Bateson's first occurrence of 'difference'. (Floridi, 2011; p. 85)

Thus, continues Floridi, according to the *diaphoric* (differential) interpretation, the general definition of datum (a single unit of data) is:

> **Dd** datum = $_{\text{def.}}$ x being distinct from y
> where the x and the y are two uninterpreted variables and the domain is left open to further interpretation.

Why are cognition and reasoning important for this thesis?

As we have seen above the simplest difference is, of course, between existence and non-existence; between what is and is not. This has very important implications for ethics, and by extension, AI ethics. As we will see further down, entities (systems) are a certain Being (existence), and one of their systemic imperatives is based on the attempt at conserving/protecting this existence from its counterpart (non-existence or destruction of Being) at least until their goals are achieved in the world. This systemic imperative is a moral imperative because it is constitutive in building moral rules and systems.

In the ontological and epistemological sense, as Niklas Luhmann would famously state, "a system *is* the difference between the system and environment" (Luhmann, 2013; p. 44). In a cognitive sense, a system uses (micro)differences in its integrated information imposed directly (direct physical changes on its structure) or indirectly (through dedicated data inputs e.g. senses) to:

a) perceive the internal and external world;

b) integrate these differences in its internal mind; and

c) choose a particular course of (in)action.

A cognitive state at a particular time granule (moment) is experienced by the system as what it is like to be the system then—the so-called *phenomenal consciousness*. However, movement from one to another and different phenomenal state enables the so-called *access consciousness* (Smith, 2017; p. 34). Consequently, access consciousness arises out of phenomenal consciousness.

Additional support for this position comes from Tononi's IIT. For example, he and his collaborators define information as one of the phenomenological axioms of experience. Each particular experience is a *different* integrated information—defined exactly by how it differs from any and all other experience. In their own words, "consciousness is informative: each experience differs in its particular way from other possible experiences. Thus, an experience of pure darkness is what it is by differing, in its particular way, from an immense number of other possible experiences. A small subset of these possible experiences includes, for example, all the frames of all possible movies" (Oizumi et al., 2014; p. 2) (see also 2.1.3 Integrated information above, where I already covered IIT in greater detail). Or, as they are keen to state, for information only “differences that make a difference within a system count” (Oizumi et al., 2014; p. 5).

In conclusion, difference is the foundational property of reality that enables cognition and reasoning. This is an important implication for the work here, and will make itself apparent in this and the next chapter.

2.1.6 Method of Levels of Abstraction

The method of levels of abstraction (LoAs) is a sophisticated method enabling determinate analysis and model-building of systems, without which it might prove an impossible task. What I will describe here will be a significantly shortened version that makes sense only for the work in this text. Readers can refer to Floridi’s *Ethics of Information* (Floridi, 2013; ch. 3) and *Philosophy of Information* (Floridi, 2011; ch. 3) where a significantly more detailed exposition on the method is included.

In essence, what the method of LoAs does is providing a way to create a *model* of a system (and its *change* through time) by creating an abstraction of it on various levels (a *level of abstraction*). This is done by including or excluding certain *observables* (which are defined as *interpreted typed variables*) from the system and their relations. The relationship between the (observables of the) model and the (parts of the) system itself should be *homomorphic*, which means that the relationship (a function) between them is pre*served (Klir, 2001; p. 95)*. This reminds strongly of the positive definition of a system in systems theory, which defines it as the set of things and set of relations (defined on the set of things) that create a model of the system (see 3.1.1.1 Positive definition of system below in this chapter).

In order to understand what all the above means, I will provide further definitions of the key elements of the method of LoAs (taken from Floridi (2011) and Floridi (2013)):

Variable	A symbol that acts as a place-holder for an unknown or changeable referent
Typed variable	Variable qualified to hold only a declared kind (type) of data
Interpreted typed variable	A typed variable that is conjoined by a statement of what feature of the system under consideration it represents
Observable	An interpreted typed variable regarding a particular system
Change	Change in the values of observables over some other metric (e.g. time) can represent change in the system (which enables conceptualizing states and transitions of the system)
Relation	Function that sits between observables (different or identical) and that describes the change (see above) between two states. It is a mathematically-described connection between those observables.
Model (of a system)	A function of the available observables

An example is in order to make the above more understandable.

Let’s take the example scenario of Piotr’s speeding car caught by police radar and issued a ticket (to its driver which we don’t know who it is exactly). The LoAs we can use here are plenty (potentially inexhaustible), but I will use some that are intuitively applicable to the scenario. We will use the following observables:

Observable	**Type**
time	a relative point in time (t_n)
car	manufacturer and model

owned by Piotr	boolean (TRUE, FALSE)
driver	person’s name
speed	km/h
speed limit (at the section of the road where the car is currently at)	km/h
speeding	boolean (TRUE, FALSE)
caught by radar	boolean (TRUE, FALSE)
issued a ticket	boolean (TRUE, FALSE)

We will start with a simple LoA (that includes **time**, **car**, **speed**), and then upgrade it so that we can see how the model upgrades.

Let’s say we have 4 points in **time**: t_0, t_1, t_2, t_3. By using them, we can describe change in the model of the system, which ought to describe change in the system itself.

time	t_0	t_1	t_2,	t_3
car	Fiat Multipla 1st generation	Fiat Multipla 1st generation	Fiat Multipla 1st generation	Fiat Multipla 1st generation
speed	80 km/h	80 km/h	65 km/h	0 km/h

We can see that with this LoA we are unable to represent the plethora of information about the system and the scenario, so all we can understand is that the same type of car is driving at 80 km/h and then cutting down in speed to 0 km/h. Our model is not suited to analyze the situation.

Now, if we add 3 more observables (**owned by Piotr**, **speed limit**, **speeding**) we can make a significantly more powerful analysis of the situation.

time	t_0	t_1	t_2,	t_3
car	Fiat Multipla 1st generation	Fiat Multipla 1st generation	Fiat Multipla 1st generation	Fiat Multipla 1st generation
speed	80 km/h	80 km/h	65 km/h	0 km/h
owned by Piotr	yes	yes	yes	yes
speed limit	90 km/h	50 km/h	50 km/h	50 km/h
speeding	no	yes	yes	no

However, we still are missing the full picture. We are analyzing a scenario of a police radar catching a speeding car and then the police issuing a ticket to the driver. There’s none of that above, so we need to add the fitting observables (**caught by radar**, **issued a ticket**).

time	t_0	t_1	t_2,	t_3
car	Fiat Multipla 1st generation	Fiat Multipla 1st generation	Fiat Multipla 1st generation	Fiat Multipla 1st generation
speed	80 km/h	80 km/h	65 km/h	0 km/h

owned by Piotr	yes	yes	yes	yes
speed limit	90 km/h	50 km/h	50 km/h	50 km/h
speeding	no	yes	yes	no
caught by radar	no	no	yes	no
issued a ticket	no	no	no	yes

And, there is one final piece of information missing. To whom was the ticked issued? So, we include the final observable (**driver**) and get a full picture of the situation (for our purposes at least).

time	t_0	t_1	t_2,	t_3
car	Fiat Multipla 1st generation	Fiat Multipla 1st generation	Fiat Multipla 1st generation	Fiat Multipla 1st generation
speed	80 km/h	80 km/h	65 km/h	0 km/h
owned by Piotr	yes	yes	yes	yes
speed limit	90 km/h	50 km/h	50 km/h	50 km/h
speeding	no	yes	yes	no
caught by radar	no	no	yes	no
issued a ticket	no	no	no	yes
driver	unknown	unknown	unknown	unknown

As we can see, by enlarging our LoA with relevant observables we can make more detailed analysis of a scenario (a system through change). This is up to a certain point, of course, since including too many observables can translate into impossibility to perform analysis because of either too little computational power (problem of complexity), or of inability to record or explicate the values of the observables (epistemic failure).

Therefore, at times, a more simplistic model of a scenario provides vastly superior capacity for study than including every possible observable there is under the sky. This implication will become very important in this work because moral reasoners are oftentimes dealing with problems of complexity that force them either not to make a decision (analysis paralysis) or refer to heuristics (see 3.4.2 Complexity below).

The only thing we don't know in our scenario, from our LoA's point of view, is—who got the ticket? This is an important implication that is worth commenting here. For example, our LoA might not include epistemically available data on who the driver was, and hence, who got issued the ticket. However, the police's LoA probably includes this information (while not having data on other observables for which we do have data about). The police's LoA might look like this:

time	t_0	t_1	t_2,	t_3
car	Fiat Multipla 1st generation (assumed)	Fiat Multipla 1st generation (assumed)	Fiat Multipla 1st generation (confirmed)	Fiat Multipla 1st generation (confirmed)
speed	unknown	unknown	65 km/h	0 km/h
owned by Piotr	yes (retrospectively assumed)	yes (retrospectively assumed)	yes	yes

speed limit	90 km/h	50 km/h	50 km/h	50 km/h
speeding	unknown	unknown	yes	no
caught by radar	no	no	yes	no
issued a ticket	no	no	no	yes
driver	Anna (retrospectively assumed)	Anna (retrospectively assumed)	Anna	Anna

Since the car enters into police's LoA only at t_2, it is natural *for them* to retrospectively assume some things that come from common sense. But they cannot even assume some other things (well, at least reasonably), like the **speed** and **speeding** in t_0 and t_1, because they don't have any data on an observable that is keen to change a lot between time points (unlike **owned by Piotr**, **car**, or the **driver**). The police LoA can also include some other observables that our LoA doesn't even contain (for example: **plate number**, **radar malfunctioning**, **officer measuring with radar**, **officer issuing a ticket**, etc). If we don't care about them it is better to not include them in our LoA at all (and thus save memory space and/or computational power). But if we do include them after all, we might simply be forced to fill their values with *unknown*.

Three comments on this implication is also in order.

Firstly, in theory, the type of an observable can be specified only to contain a closed set of values. But it would be best to simply automatically append the values of *unknown* and maybe even *undefined* to the possible kind, so that it automatically covers for any kind of situation in regards of value of an observable (and avoid simple programming errors). The value of *unknown* can be automatically assigned to an observable where no value input has been provided, either manually or through a process (e.g. extracted from a database, inputted by an algorithm, a device or a sensor, and similar).

Secondly, we can notice that LoAs are practically point of views. They are an epistemological method that enables model building (and a model *is* a certain kind of a perspective). It is important, then, to not mistake the model of a system with the system itself. However, since the aim is to model a system for a particular purpose, the best attempt should be made to make the model follow the actual system as closely as possible i.e. the model ought to aim to be in a homomorphic relation with the system. This typically means including or excluding observables, reformatting them (modifying their types to fit other kinds of data), and of course, making the best data-gathering possible with the most suitable instruments.

One final point to make. All of the above analysis can of course also be represented in a graphical format. Also, other observables such as the movement of the car, its location, the location of the radar … can all be represented in a visual space, or even on a simple multidimensional Cartesian plane. In fact, graphical simulations of processes and systems are exactly this, and they help convey the process in a format more acceptable by our evolutionary-imposed interface—our visual system. I will be using tabular representations further down in 4 Towards Ethics of Systems, and in Chapter IV. Applying Ethics of Systems Framework to AI moral scenarios (the Ethics).

2.1.6.1 Is the LoA method essentially a heuristic?

The method of LoA sits somewhere between a heuristic and a formal method. Although its proponents might disagree because *nominally* it is a complete, well-formed, and coherent method, *in practice* it is not. Let me explain.

In essence, all our methods we use for research are heuristics of a sort. The reason being: complexity. We attempt to gather as much relevant data as we can about a phenomenon, organize it in a meaningful manner, and extract information that tells us more about the subject of study. However, because of inherent limitations of our data gathering and processing capacities (Harshman, 2016; p. 8), even the best methods are necessarily limited. We can never establish the ultimate answer to a research question. The same goes for the method of LoAs.

For instance, even though we improve the methodology on spatial measurement we might never be able to measure a smaller distance than the Planck length. That would mean that we might also never be able to spot what is going on in between those lengths (if anything). Those phenomena, if they exist, may forever remain inaccessible for our perception, measurement, and modeling.

Similarly for the method of LoAs. We can make successful data gathering and processing of simple Trolley-type scenarios with certain observables. But modeling the morality of a crowd with the same observables becomes exponentially more difficult, of a nation even more, and of the whole planet—probably impossible. We would be forced to exclude certain observables (or merely assume their values) and other properties from our LoA so that we make the model computationally-feasible. For example, we will most certainly not attempt to model the interactions of people's atoms in a traffic Trolley scenario (unless it is really, *really* essential for the model).

This exclusion (of certain observables and properties e.g. relations) weakens the model building capacity and *effectively* turns the method into a heuristical one, because we choose to disregard some kind of data to be able to make a model or a conclusion in a practical time-frame. This is the very definition of a heuristic. Just like the hyena being chased by a lion chooses to disregard all the zebras, trees, and airplanes around it and focuses on what's important, we also exclude things that we deem not *that* important (although they very well may be).

Similarly, Floridi has this to say on the method:

> The method of abstraction is ideally suited to the study of systems so complex that they are best understood stepwise, by their gradual disclosure at increasingly fine or alternative levels of abstraction (Floridi, 2011; p. 63).

Therefore, theoretically, the method of LoAs is sound and complete—but practically it probably will always remain a heuristic (just like any other method). We are using abstraction to handle complexity, or, in the words of E. W. Dijkstra, "... the purpose of abstracting is not to be vague, but to create a new semantic level in which one can be absolutely precise" (Van Leeuwen, 2014). The upper limit of computational capacity for now is the Bremermann's Limit (see 3.4.2 Complexity).

But this need not be a disaster for science and research, especially in ethics. We can recall that *all* moral reasoners, whether entities embedded in a moral scenario, or outside spectators and analyzers of it, are inherently limited and thus always fall back to heuristics when complexity exceeds their reasoning capacity. This is why moral reasoning which is distributed on the many components of a system (e.g. participants in a debate, employees of an organization, and voters in a democratic political organization) has the potential to provide better conclusions and suggestions for (in)action compared to single individuals (but not necessarily, of course).

Alas, this probably means that (moral) bias will always remain part of very complex moral scenarios even with our improving computational machinery and algorithms (see Chapter V. Discussion section 2.3.1. Substantial (ethical) implications).

2.1.6.2 The systemic LoA and its place in AI ethics

Floridi mentions that the informational LoA that ethics of information (IE) uses offers advantages over a biological one, or even others i.e. anthropocentric, biocentric, ontocentric LoAs (Floridi, 2013; p. 35). Even though the difference is small, what I argue for here is that the *systemic* LoA offers even greater (if at times negligent) advantage over the informational and all other LoAs. This is why here the systemic LoA is taken as the interface through which moral scenarios are understood, modeled and managed.

It has to be remembered, though, that other LoAs—or at least their elements (observables and relations)—have to be taken into consideration and thus appended to the systemic LoA in particular situations, where appropriate. What is going on here is not that we are ‘upgrading’ the systemic LoA, but that we are amending the model we have of the system itself by including (or even excluding) observables and relations that previously we did not or could not take into account. More about this in the text below.

But what is the systemic LoA?

The systemic LoA is the one that includes observables that are the subject of study in systems science. You might recall from above that LoAs and models are in essence points of view or perspectives. There are two points of view that are applicable to any system: what systems *are* and what they *do* (see 3.3 Systemic Being and 3.4 Systemic deliberation below). Let’s call them the *existential* and the *functional* view[18]. Therefore our systemic LoA has to include observables that describe both these views.

Since according to systems theory’s positive definition of a system (see 3.1.1 A system) a system is a set S comprised of two sets, the set of things T and the set of relations R, I will include these as observables. However, this only covers the statics of a system. The dynamics is contained in its set of goals, so I will also include systemic goals as an observable. Goals make no sense without data and information about the current state of the system, so personal feedback observable (a set of states) also ought be included. And finally, an observable that identifies a system (an identifier) ought be included. It would identify a particular system (an *instance*), a *class*, or both. With the aforementioned I have covered the existential view, and for many purposes this view is enough.

Nevertheless, if and when we want to create a dynamic model of what systems do, and thus to account for the functional view, we ought also include observables that are fit to describe interactions with their environment, which is, everything outside their boundary. These observables are different according to the scenario we are trying to model, so they cannot be firmly set up front. But they do include things such as resources, other systems, position, danger, movement, speed, outside data and/or information, and other observables that belong or pertain (as in the case of position) primarily to their environment.

With all the above being said, I suggest the systemic LoA to include the following observables and types:

[18] They are, essentially, separate LoAs.

Table 4: The systemic level of abstraction

View (LoA)	Observable	Type
Existential	ID	Name of system (instance and/or class)
	Set (things)	Components of the system (physical, informational, and other)
	Set (relations)	Relations between the components
	Set (goals)	Goals of a system (implicit or explicit; imperatives and regular goals)
	Set (states)	States of: goals (achievement, ordering), available resources, healthiness, ...
Functional	Set (resources, data, information, position ...)	*[to be defined when appropriate]*

With the systemic LoA we can perform the fundamental model building of systems and scenarios, including moral scenarios, which will be of great importance for this work as we will see in the further text. Of course, other observables will be appended to the systemic LoA where appropriate i.e. identifiers of moral entity (agent, patient, neutral), points in time, descriptors of future (potentiality) and past (past states through points in time), etc.

Why is the method of LoA important for this thesis?

As Floridi himself states, "A theory comprises at least a LoA and a model. The LoA allows the theory to analyse a given system and to elaborate a model that identifies some properties of the system at the chosen LoA" (Floridi, 2013; p. 34).

It is a method that enables explication of elements of the objects of study e.g. of moral scenarios where AI entities are involved. With this it aids analysis of their different components and their relations and interactions, which can then be merged into a complete, holistic picture on which attempts at establishing a set of solutions can be made. As we can see, it is a very applicable tool in the field of AI ethics, especially when used with the appropriate fundamental LoA (which I argue above that is the systemic LoA). Thus Floridi:

> The method clarifies implicit assumptions, facilitates comparisons, enhances rigour, and hence promotes the resolution of possible conceptual confusions. If carefully applied, the method confers remarkable advantages in terms of consistency and clarity. Too often, philosophical debates seem to be caused by a misconception of the LoA at which the questions should be addressed. This is not to say that the method represents a panacea. Disagreement is often not based on confusion. Indeed, informed and reasonable disagreement is precisely what characterizes philosophical questions, which remain intrinsically open to debate. But the chances of resolving or overcoming it, or at least of identifying a disagreement as irreducible, may be enhanced if one is first of all careful about specifying what sort of observables are at stake and what goals are orienting their choice, and therefore what questions it is meaningful to ask in the first place (Floridi, 2013; p. 52).

I will be using the method of LoAs to explicate moral scenarios and their components further down in this work, primarily with the systemic LoA, upgraded where appropriate.

2.2 The *ethics* of information

The philosophy of information is a fairly straightforward study—it has its own subject, methods, publications, best practices, and so forth. However, is it natural to assume that there would be an *ethics* of information stemming from it? Probably not on first sight.

However, as Floridi is keen to comment, there is currently undergoing a process of transformation of our philosophical anthropology and metaphysical outlook. This transformation is caused by the so-called Fourth Revolution, brought about the widespread introduction of Information and Communication Technologies (ICTs) in our environment. This also affects our moral standings, by providing novel or reinterpreted moral issues that did not appear before (e.g. privacy, moral action, the infosphere, etc.). Therefore, there is a need to develop a new ethics, *ethics of information* (IE).

IE, according to Floridi, is under the threat to remain a *microethics*, dealing with particular domains and issues, and being treated as a simple extension of other *macroethics*. However, he argues that IE can and ought be developed as a fully-fledged macroethics (Floridi, 2013; ch. 2.5.). IE as a macroethics thus would be able to "clarify and solve the ethical challenges arising in the infosphere" (Floridi, 2013; p. 19) (see also 2.1.2 The infosphere).

In the interest of space, there are some key elements of IE that I will explore here: informational Good and Evil, IE's four ethical principles, moral entities, moral scenario, and the model of moral action.

2.2.1 (Informational) Good and Evil

There is no complete ethical study without a study on Good (and/or Right) and Evil (and/or Bad and Wrong). In general, the Good/Right is what we strive for with our morally positive behavior, moral principles and action; while Evil/Bad is what we strive away from (except in some cases of a *necessary Evil*).

Floridi's IE has its own conceptions of Good and Evil that I am going to cover here.

2.2.1.1 The Good

Floridi does not seem to provide an explicit, straightforward definition of the Good[19]. However, he does provide the four ethical principles of IE (see 2.2.2 The four ethical principles of IE below) for which he comments that they are listed in their increasing moral value. The fourth principle has the highest moral value and thus can potentially be taken as a definition for the Good. So here follows my own construction of the Good for Floridi's IE:

> Good $=_{\text{def.}}$ "the flourishing of informational entities as well as of the whole infosphere ought to be promoted by preserving, cultivating, and enriching their well-being" (Floridi, 2013; p. 71).

IE is an environmental ethics (recall 4.1.2.3 Environmental ethics and deep ecology in Chapter II. Literature review)—that is, it deals with ethics of the environment (the infosphere). On IE, Taddeo further comments that **the Good is represented in the flourishing of the infosphere and all its components** (Taddeo, 2016; p. 369) – the informational entities in it. Hence, Taddeo:

> „Informational good, as the opposite of the metaphysical entropy, is any form of flourishing of informational entities and of the infosphere. IE endorses an environmental approach, which rests on moves (a) and (b) described in the fourth section. The ultimate patient, whose well-being ought to be at the center of the moral concern of any agent, is the informational environment. This is a key point of IE, [...]". (Taddeo, 2016; p. 369).

[19] He implies it with phrases such as uniformity of Being and intrinsic value of Being.

That is to say, flourishing is the final target that moral entities ought to strive for. Flourishing is a contentious term (just like Good and Evil), so it bears some explanation here. In Floridi's IE it means "to improve and enrich its existence and essence" (Floridi, 2013; p. 84). Those that are knowledgeable in virtue ethics (especially Aristotle's *Nichomachean Ethics* and with the Stoics) would immediately notice the resemblance with their concept of *eudaimonia*, which is most consistently translated into *flourishing* in English.

Flourishing according to virtue ethics is development of character through development of virtues to the maximum of one's potential. It is living the good life through excellence (virtue, *aretē*) and practical and ethical wisdom (*phronesis*). In short it is achieving the best possible way of moral and personal living by personal development.

However, most variants of virtue ethics belongs to the 'classic' ethical theories (the other being deontology and consequentialism) who are agent-oriented (although ethics of care is sometimes taken as patient-oriented virtue ethics). In contrast, IE is part of environmental ethics and thus patient-oriented (Floridi, 2013; p. 65). Is this difference irreconcilable? I, and Floridi also, would argue—no. Even though IE is patient-oriented, it still takes into consideration moral agents. Moral agents are expected to be virtuous and help the flourishing of moral patients and the whole infosphere. Additionally, in the same moral scenario the moral agent can in parallel be a, or the, moral patient (and hence pursue its own flourishing).

Two keywords here strike the curiosity—existence and essence. *Nothing* cannot flourish, only *something* can. All (informational) entities hence have a particular and unique existence that can either flourish, conserve, or languish. Therefore, IE's starting point is the recognition that "any expression of Being (any part of the infosphere) has an intrinsic worthiness. So any informational entity is to be recognized as the centre of a minimal moral claim, which deserves recognition in virtue of its presence in the infosphere and should help to regulate the implementation of any information process involving it, at least prima facie and overridably" (Floridi, 2013; p. 84).

Additionally, any informational entity has "a Spinozian right to persist in its own state, and a constructionist right to flourish (...), i.e. to improve and enrich its existence and essence"[20] (Floridi, 2013; p. 84). **This is what is Good.**

We can see here that Being has two perspectives—static and dynamic. This is recognized also in systems theory (see 3.3 Systemic Being below), and will be part of *ethics of systems* (see 4 Towards Ethics of Systems in the further text).

Non-monotonicity and resilience of the Good

One further commentary is in order here. In contrast to Evil (see immediately below), the Good is non-monotonic (Floridi, 2013; p. 72). There's no one, final Good that can be achieved. Furthermore, even some other good actions can sometimes be reinterpreted as evil, depending on their effects. Goodness "can, in principle, turn out to be less morally good and sometimes even morally wrong unintentionally, depending on how things develop, that is, on what new state the infosphere enters into, as a consequence of the process in question" (Floridi, 2013; p. 72).

However, goodness is resilient, in two senses. In a *fault-tolerant* sense, goodness "has the ability to keep the level of entropy within the infosphere steady, despite the occurrence of a number of negative processes affecting it" (Floridi, 2013; p. 73). In an *error-recovery* sense, goodness has to some extent "the ability to

[20] These two outrightly relate to the two moral imperatives of systems, Conservation of Personal Continuum and Achievement of Personal Goals (see 4.3.2 The moral imperatives).

resume or restore the previous entropic state of the infosphere, erasing or compensating any new entropy that may have been generated by processes affecting it” (Floridi, 2013; p. 73).

2.2.1.2 The Evil

In IE, Evil, in contrast to the Good, is *monotonic* (Floridi, 2013; p. 72), that is to say, Evil is always evil and cannot be reinterpreted as Good. Of course, there are levels of evil as not every (in)action is equally destructive towards Being and the infosphere (because not every action introduces an equal, or equally-important, amount of entropy).

But what is Evil?

Since Being and its flourishing is the Good, the opposite should be Evil. Opposed to Being is movement towards non-Being (the so-called nonsubstantialism: the deflationary theory of evil (Floridi, 2013; p. 183)), and since any Being is a certain kind of dynamic order and structure in the infosphere, the opposite to that order is *metaphysical entropy*. Metaphysical entropy is what destructs Being and the infosphere by damaging the order and structure that exists and thus disabling their functioning and flourishing.

But we have to be careful not to err by believing that metaphysical entropy *does* and *is* **something**. Metaphysical entropy is the exact opposite of Being, so it cannot *be something*. Similarly, it does not do something, but is the opposite (or incapacitator) of doing. This might be counter-intuitive at first. Sometimes this might be interpreted as if evil does not exist (in the sense of Being, like an entity). But evil is the destructive *effect* on Being, and since only actions cause effects, only actions can be evil—not things (entities) (Floridi, 2013; p. 184). “Evil exists not absolutely, per se, but in terms of damaging actions and damaged patients. The fact that its existence is parasitic does not mean that it is fictitious” (Floridi, 2013; p. 184).

> Floridi refers to informational evil as to metaphysical entropy. […] Metaphysical entropy has a very specific meaning, for it refers to any form of destruction of information and as such of Being. It indicates the opposite of semantic and ontic information […], as such metaphysical entropy refers to the decay, the corruption, of content of the infosphere and of the entities inhabiting it, and hence it is a form of impoverishment of Being. Insofar as Being is co-referential to the infosphere, metaphysical entropy is analogous to the metaphysical concept of nothingness. Corrupting a file or damaging a piece of art, violating someone’s privacy and killing a living being are all examples of metaphysical entropy. (Taddeo, 2016; p. 369).

So the time has come to provide a definition for Evil:

> Evil = $_{\text{def.}}$ metaphysical entropy; movement towards destruction of Being.

But since only actions can be evil, a definition of evil action is far more usable for IE:

> Evil action = $_{\text{def.}}$ one or more negative messages, initiated by A, that brings about a transformation of states that (can) damage P’s well-being severely and unnecessarily; or more briefly, any patient-unfriendly message. (Floridi, 2013; p. 183)

This notion of metaphysical entropy as destruction of order and Being is also shared by Norbert Wiener. According to Bynum, “The second law of thermodynamics applies to every physical change in the universe, and Wiener realized that an increase of “entropy” amounts to a loss of physical information” (Bynum, 2016; p. 205). Similarly, Wiener in his keynote speech of 1946 at the New York Academy of Sciences declared that:

> “Entropy here appears as the negative of the amount of information contained in the message… In fact, it is not surprising that entropy and information are negatives of one another. Information measures order and entropy measures disorder” (Wiener, 1946, quoted in Conway and Siegelman, 2005, p. 164) (Bynum, 2016; p. 204).

In IE there are 3 types of Evil: natural evil (caused by natural phenomena and entities, such as rivers, earthquakes and similar), moral evil (caused by morally-responsible agents i.e. humans), and artificial evil (caused by artificial entities). I am simply mentioning this distinction for when there would be need to make it in the further text.

This deflationary interpretation of evil will be a very important component of the *ethics of systems* that I am developing at the end of this chapter (see 4.2.2 The Bad and the Evil).

2.2.2 The four ethical principles of IE

In order for a moral entity to be morally-principled, it ought to include in its own moral theory (ethic) several rules of ideal conduct or aims that are to be pursued to the best of the abilities. These rules and aims are the ethical principles of that theory. More complex decision-making should be possible to be derived from them. In essence, ethical principles are the formal aspects of a theory.

IE includes four of them. Predictably, they are based around flourishing (Good) and entropy (Evil) as the opposite poles, are listed in order of increasing moral value, and are formulated in a patient-oriented version. Hence,

> 0 entropy ought not to be caused in the infosphere (null law)
> 1 entropy ought to be prevented in the infosphere
> 2 entropy ought to be removed from the infosphere
> 3 the flourishing of informational entities as well as of the whole infosphere ought to be promoted by preserving, cultivating, and enriching their well-being (Floridi, 2013; p. 71)

When a moral process (e.g. action) is failing to satisfy a principle in a decreasing order, it is increasingly disapprovable and its agent-source increasingly blameworthy. On the contrary, if a process satisfies the null principle and at least one other law, it is approvable and its agent-source praiseworthy. This approvability and praiseworthiness increases as the number of satisfied principles increases.

Finally, a process that only satisfies the null principle does not change the moral state of matters in the infosphere. Such a process is morally irrelevant, insignificant, or negligible (Floridi, 2013; p. 71).

2.2.3 Moral entities

Moral entities are entities that are relevant from the aspect of morality. This is because they are either the producers of moral actions, or the receivers of their effects. IE recognizes two types of moral entities: agents and patients.

2.2.3.1 MORAL AGENTS

Moral agents are the producers of moral action, or in IE-theoretic terms, senders of moral messages.

> Agent =$_{\text{def.}}$ a system, situated within and a part of an environment, which initiates a transformation, produces an effect, or exerts power on it over time. (Floridi, 2013; p. 140)

Traditional theories focus on the actions of moral agents and what they ought (not) do. IE's ethical principles (see immediately above) can, in principle, easily be reformulated in an agent-oriented manner.

2.2.3.2 MORAL PATIENTS

In contrast to agents, moral patients are the entities that are receivers of the moral message i.e. moral action is being performed *on them*.

> Patient = def. system that is (at least initially) acted on or responds to a transformation, production of an effect, or exertion of power over time. (Floridi, 2013; p. 187)

Contrary to traditional theories, some newcomers (i.e. environmental ethics, ethics of care, and now IE) are patient-oriented. They aim to stipulate what ought (not) be done to patients, and they thus formulate principles in this manner.

2.2.4 Moral action

The model of moral action is, basically, ∃*A* ∃*P* M(*A*, *P*) (Floridi, 2013; p. 181). That is, there exist an agent and a patient; the agent sends a moral message to the patient. However, in IE there is more to moral action. Altogether, the elements of moral actions are the following (Floridi, 2013; p. 108):

1. *A* = Alice the moral agent
2. *P* = Peter the moral patient
3. *M* = moral action, constructed as an interactive information process
4. *shell* = Alice's (and Peter's) personal world of information
5. *factual information* = information about the moral situation
6. *envelope* = the moral situation
7. *infosphere* = the general environment

I will also copy the illustration from Floridi's book (Floridi, 2013; p. 108), because I will later modify it to suit ethics of systems' concept of a *moral scenario* (see 4.4 The moral scenario).

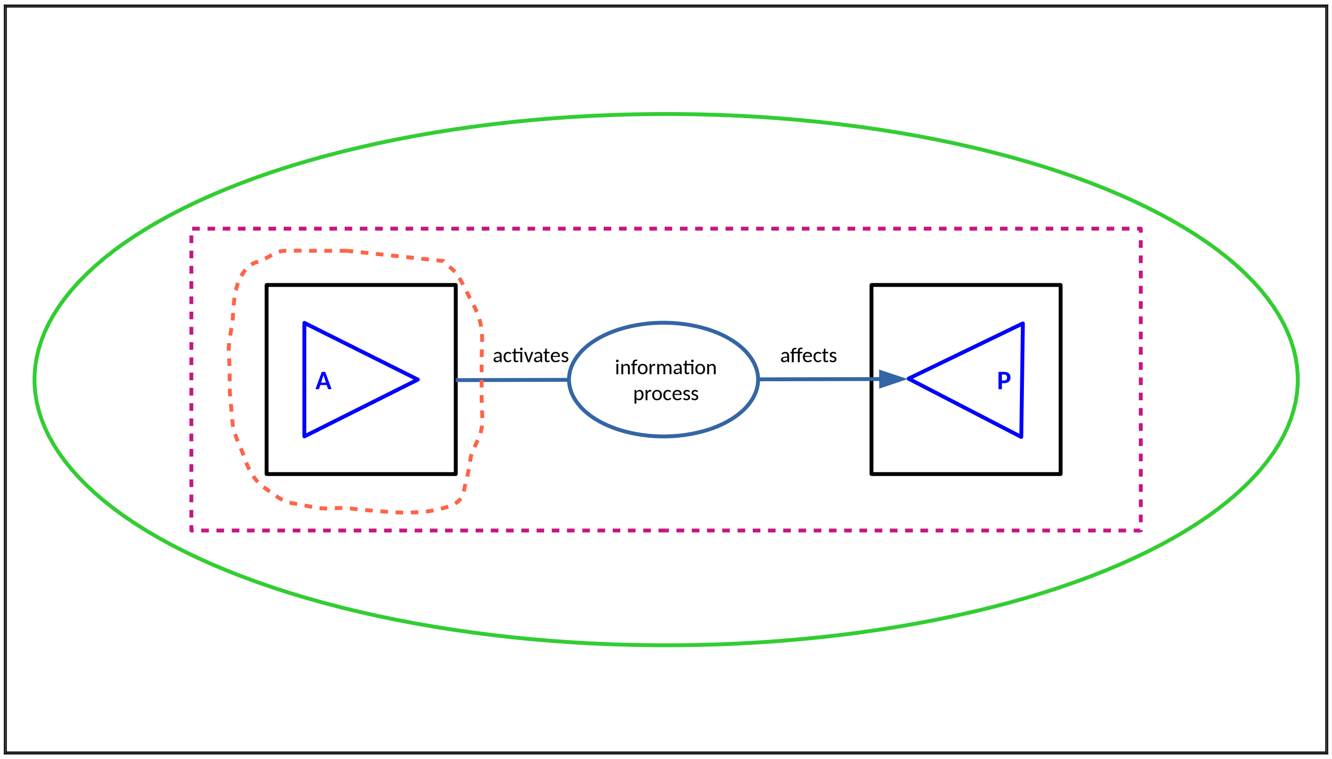

Illustration 3: Moral action in IE. Components: agent and patient, moral action, shell, factual information, envelope, the infosphere.

2.3 Final commentary on IE

As we have seen above, ethics of information has several methodological and theoretical developments that prove very useful for ethics, and AI ethics in particular. I will use some of its tools and findings to aid systems theory and derive *ethics of systems* at the end of this chapter.

IE is based on the informational LoA. Systems theory is based on a conceptually more fundamental LoA, the systemic one (see 2.1.6.2 The systemic LoA and its place in AI ethics above). However, I will be upgrading the systemic LoA with the informational LoA (since, naturally, oftentimes information is being sent between systems), and with observables from other LoAs where it fits.

3 Systems theory

It is my intuition that systems theory can contribute significantly to AI ethics since it provides a concise, holistic, and scientifically rigorous approach to systems—which can describe and include *any* and *all* types of entities *and their interrelations*, such as people, societies, ecosystems, planets, and artificial entities (i.e. companies, AI entities, computer systems, and similar).

Systems theory provides a necessary low-level layer of analysis and synthesis which is applicable to all the above *and their relationships*. Having into mind that ethics is, at least in a significant part, focused on relationships between entities, often of different types and with differing goals and states, I chose to include the basic tenets of systems theory and to attempt to apply its approach to the subject matter of this work—AI ethics.

3.1 Systems

3.1.1 A system

I already mentioned in Chapter II. Literature review that in systems science there are such systemic properties such as *systemhood*, *thinghood*, and *setness*; and they are somehow applicable to the study of systems. But to get into them we first need to dive into the substance and definition of a system.

Although the term *system* is widely used in scientific and general literature, and elsewhere, rarely there is a definition of what exactly a system is. What a system is, is assumed to be understood by all participants in the discourse. However, even a cursory inquiry can provide insight into how elusive this term is, and how it is often used in inappropriate manner. To avoid this and similar misunderstandings, we ought to offer a working definition of what a system is.

Typically, there are two approaches that can be used to define systems, which I describe as a *positive* and a *negative* approach. The qualifier of *positive* is used in this respect to denote a definition of what a system *is*. The *negative* qualifier, on the contrary, aims to describe what a system *is not*. Both of these approaches have their advantages and disadvantages, so I will include them both here.

3.1.1.1 Positive definition of system

The typical, 'common-sense' definition used in a positive sense is that a system is

> "'a set or arrangement of things so related or connected as to form a unity or organic whole' (Webster's New World Dictionary)" (Klir, 2001).

However, this common-sense definition of what a system is not rigorous enough for scientific inquiry. Thus, Klir suggests a more analytical formulation of the common-sense definition:

> “It follows from this common-sense definition that the term "system" stands, in general, for a set of some things and a relation among the things” (Klir, 2001) (see also (Bertalanffy, 1969; p. 55)).

and with a symbolic representation:

> **S** = (**T**, **R**)
>
> where
> **S** denotes a *system*,
> **T** denotes *a set of things* distinguished within **S**, and
> **R** denotes *a relation* or *a set of relations – defined on **T*** (Klir, 2001).

Let’s agree to label this property as a system’s ***is*-ness**.

The *thinghood* and *systemhood* properties of S reside in T and **R** respectively. What characterizes a particular ***set*** of things (i.e. a *collection* of particular books) is the property of ***setness***. What characterizes the ***things*** themselves in that particular set (i.e. which particular books are there), is the property of ***thinghood***. And, finally, what characterizes the particular ***relations*** between the *things* in the particular *set* (i.e. how they are organized) is the ***systemhood*** property. More about these properties in the following subsection.

S, **T** and **R** can be used as symbolic representations, and analyzed and manipulated using mathematical, logical and argumentation tools. This work will only include basic level of such analysis, manipulation and representation which will be enough for the purpose of the study.

3.1.1.2 Negative definition of system

However, as mentioned above, there is another approach in defining what a system is not, which I described as the negative, or *differential*, approach. Let’s agree to label this property as system’s ***is*-not-ness**. At first glance, such an approach might be questionable, since why would a rigorous study need a definition of what something *is not*, as this cannot directly be used to perform study on the subject.

However, what a system *isn’t* is the exact opposite of what a system *is*. This can provide us with a powerful, Popper- and Occam-inspired tool[21] to determine the boundaries of a system (or class of systems) and thus to find out what exactly is a particular system. Thus, we may use Niklas Luhmann’s notion of *system as a difference*. In his own words,

> “a system *is* the difference between system and environment” [emphasis original] (Luhmann, 2013; p. 44)

Every system exists in its *environment*. The system typically interacts with its environment; but the system is *not* the environment. (A part of) the environment can become part of the system (can integrate with it), but then it ceases to be a part of the environment anymore (when looking at the system per se).

In a certain scenario that contains a particular system and its environment, if we start to progressively eliminate everything, up until a certain point we can eliminate a lot—and the system will remain the same. Further elimination past this point will change or destroy the system itself. This ‘point’ is the *boundary* between the system and its environment. Similarly, Oizumi et al. (2014; p. 12): “Since the set is reducible without any loss, it does not exist intrinsically – it can only be treated as “one” system from the extrinsic perspective of an observer”.

[21] Or, as would the oft cited quote, but probably erroneously attributed to DaVinci, states: “Simplicity is the ultimate sophistication”. However, it seems that these words were firstly attested as used by the playwright Clare Boothe Luce, and in the well known form above, by Leonard Thiessen (O'Toole, 2015).

We will see further down the text that what a system is and is not will be very important from an ethical point of view. Some initial reasons mentioned here are that a system's existence and goals naturally 'emerge' out of its essence (its *is*-ness), and the system is in a complex interplay with its environment (its *is-not*-ness). The system can exert its will on the environment in the pursuit of its systemic imperatives; but also the environment can exert its power on the system, at times changing it or even destroying it.

3.1.2 The properties of thinghood, setness, and systemhood

We can take a particular set of books and throw them at random. Although the set will remain that same particular set, the books will not be organized in any particular manner, thus they won't represent a system. However, if we apply certain rules to order them on a shelf (i.e. by alphabetic order), then a relation is established between the books (the *things*) of that particular set. That relation is *alphabetic order(ing)*. Since there is a certain relation established between the books of the particular set, the relation together with the things become a system.

3.1.2.1 Systemhood

To explain the systemhood property, we can further imagine extracting that relation (alphabetic order(ing)) in some form of a representation, i.e. a symbolic description, or a mental concept. Then, we can take that relation and apply it to a *different set of things*. Let's say we apply it to a set of song files in a particular folder on our computer. We order them in an alphabetic manner, and the relation between them will be exactly the same — alphabetic order(ing).

However, the system will be different, because the set of *things* (its *thinghood*) is different. That is, a different system will *emerge* out of the different components. In the first case we have books, in the second—digital song files; but the relation is of the same type. Thus, we can extract particular systemic properties (systemhood, thinghood, setness) that determine the particular systems we have under our study.

A system comprised from a set of things **T**, and a set of relations $\mathbf{S_1}$, will be a different system than a system comprised from the same set of things **T**, but of a different set of relations $\mathbf{S_2}$, ($\mathbf{S_1 \neq S_2}$). Similarly, system comprised from a set of things $\mathbf{T_1}$, and a set of relations **S**, will be a different system than a system comprised from a different set of things $\mathbf{T_2}$, ($\mathbf{T_1 \neq T_2}$), even with a same set of relations **S**.

Hierarchy of systems

Another renowned author in the field of systems science, Mihajlo Mesarović, has introduced the notion of *complex system*. By his definition, a complex system is

> "… a system whose objects are systems in their own right. That is, a complex system is defined by a relation that describes interactions among a set of simple systems" (Klir, 2001; p. 60).

This opens up the discussion on the difference between collective and individual systems, and if there is, truly, such a difference in reality. The real question is something along the line of: **is there a difference between the whole and the parts, and where it is?**

The answer may be hiding itself in (*the*) context. When a system is the sole focus of the study attention (i.e. implicit or explicit model-building), it 'becomes' the whole. But if it is studied as a component of another system, it will 'become' a part of another system. This status of *whole* or *part* is not absolute, as Klir is apt to point out:

> "The status of a system as either a whole (an overall system) or a part (a subsystem) is, of course, not absolute. The same system may be viewed in one context as a whole and in another context as a part.

> We may say, more poetically, that a part is a whole in a role (in one context), and a whole is a part in a role (in another context). This duality makes it possible to represent systems hierarchically in the sense that a system conceived as a whole may consist of interconnected parts that themselves are systems, and each of these parts may again consist of interconnected parts that are systems, etc., until some primitive parts are reached that do not qualify as systems" (Klir, 2001; p. 42).

However, we ought be careful as this strongly reminds of systemic constructivism i.e. the view that systems do not exist in nature, but we 'construct' them in our minds as models of reality. It is easy to assume that since we can choose the view and thus suddenly change from seeing the system as a whole or a part, that might mean that there is not such difference at all; and from here to jump to the conclusion that there is no difference between a system and its environment, and hence to the conclusion systems do not exist in nature at all. I do not support this argument in this form. For more on the subject please see 3.2 Constructivist realism.

Interestingly enough, the way we are looking at the same system and seeing a 'different' representation i.e. creating a different *model* of the system (a whole, a part, a symbiosis between those two) depends on our *level of abstraction* (LoA). It is what Luciano Floridi explains as what "… makes possible a determinate analysis of the system" (Floridi, 2013; p. 31). Assuming we have complete, or at least high-quality access to data/information about a system, we choose to abstract or even disregard some of these found on a certain level of abstraction that we are not interested in; and accept others, on the level of abstraction we *are* interested in, to be able to pursue our analysis of the system.

Therefore, the following is important to note. If we explicitly include components and relations of the system in our level of abstraction, we will be looking at them as *parts*, and the whole system as *whole*. If we don't include any of those in the LoA, we will just see the system as a *whole*. Now, if we include the same system as a component in another system in our LoA, the first system will become (i.e. be recognized as) a *part*. We will see that levels of abstraction as a concept, and the method itself, will comprise a significant part of the multidisciplinary approach employed in this work further down the road. I go more deeply into discussion on it in this chapter's section 2.1.6 Method of Levels of Abstraction).

Thus, what I oftentimes refer to as 'collective[22] systems' are, in fact, Mesarović's notion of a complex system. Buildings, planets, societies, economies, cars, moral systems, documents—can all be considered complex, or collective, systems. For example, societies are made out of people, infrastructure, institutions, and other systems; people are made out of various organs in specific and varying relations; organs are made out of various tissues in different relations, who are made out of various molecules in different relations, who are made out of various atoms in different relations, who are made out of various subatomic particles in different relations … (see Illustration 4 below).

[22] Collective, as in, built up from a collection of other systems.

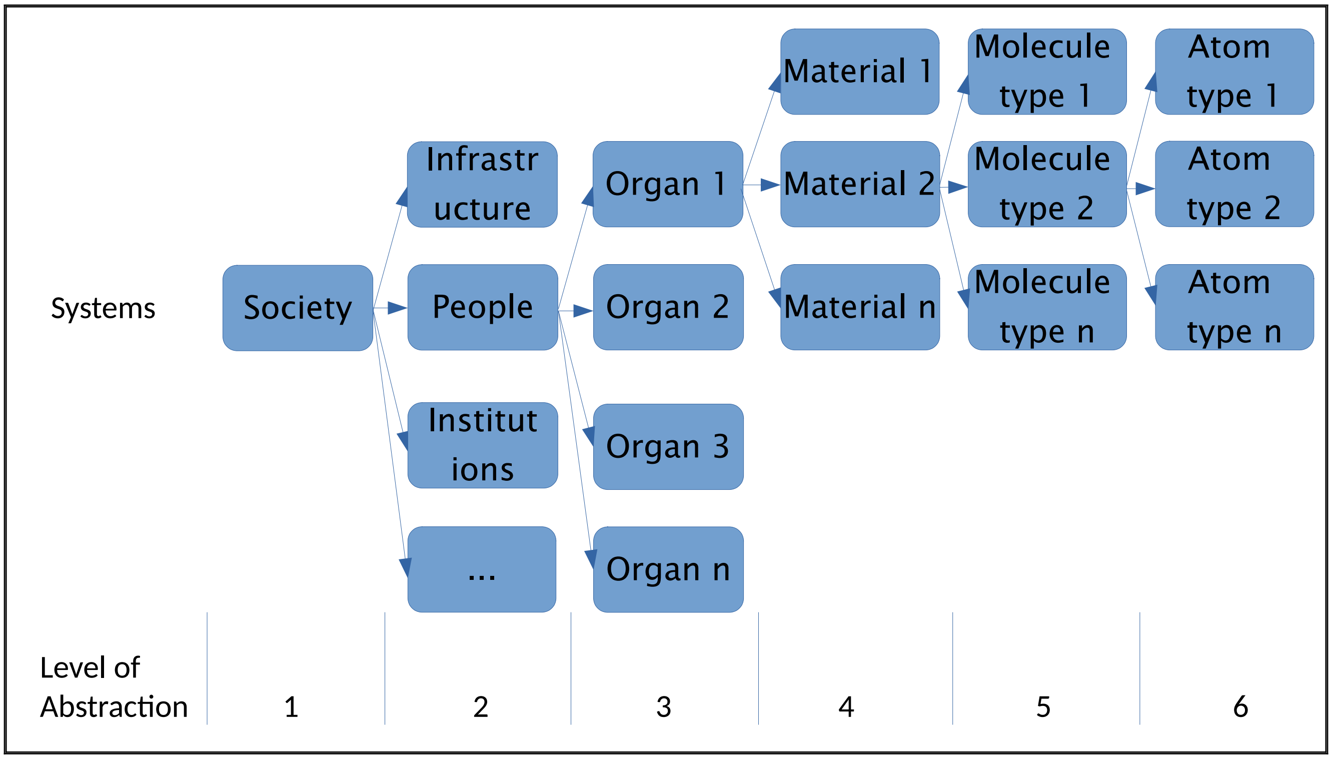

Illustration 4: An example of systems hierarchy (note: the numbers and the boundaries of the LoAs are arbitrarily chosen)

The question is **whether this goes ad infinitum, or if there is such thing as 'fundamental' primitive parts that build the first layer of simple systems?** Klir seems to assume such parts exist. However, I don't think current science has the capacity to answer this question with the necessary degree of certainty. Regardless, this does not undercut the argument of systems hierarchy that I just described.

Systemic unity through integration of its components into a unified whole is a phenomenon that also seems to be translated into consciousness itself. As I already purported in 2.1.3 Integrated information and 2.1.4 Panpsychism above, and as I will revisit the subject below, if the nature of reality (including its physical aspect) is panpsychist (i.e. there is a mind-like property in everything), an integration of a set of things into a system would automatically translates to integration of (parts or contributions of) the separate minds of the components into a new, unified mind (consciousness). This is a kind of the so-called *homuncular functionalism*, one of the theories used in neuroscience and theory of mind to explain how the brain derives the higher-level mind by the coordinated activity of smaller and more 'stupid' components ('homunculi') (Smith, 2017; para. 4.2.3.), which is, in essence, systemic emergence (see below in *3.1.3 Emergence*).

Therefore, thinking about a brain in the pure sense of its mechanics (e.g. the exchange of neurotransmitters, the firing of neurons, the uptake of food, the expulsion of metabolic byproducts and waste, etc.) will never derive the higher-level emergent result—mind and consciousness. Vice versa, thinking about the mechanics of the brain in a purely psychological concepts (i.e. traits, ego, emotions, episodic memory, qualia, etc.) will never derive to the lower-level physical and chemical mechanics. We would be doing what Ryle Gillbert would describe as a 'categorical mistake' (Ryle, 2009; Ch. 1 para. 2 and 3); and we would be mistaking the correct level of abstraction.

A repeated note here would be in order: as I already mentioned in 2.1.4 Panpsychism, the unified mind may not integrate the totality of its components' minds. Thus, the components can retain some separation in consciousness (i.e. a sense of self that cannot directly access the experience of the higher-level mind), even though at the same time they are united in a higher-level consciousness.

3.1.2.2 Thinghood

The *thinghood* of a system is contrasted to its *systemhood* by the following property. In general, whenever we are working with a system we recognize certain parts of what we are working with (the model) as its *primitives*, and some other parts as *relations* between the primitives. These primitives are otherwise called *things*. Thus, the thinghood of a particular system resides in its primitives (Klir, 2001; p. 23).

Systems science is focused on studying the systemhood of systems. However, it pays to bear in mind that when dealing with *particular* systems (entities), the thinghood is also an important property that we cannot disregard. For example, when we are dealing with the particularities of a moral situation where human entity Perunika is involved, the things that comprise that entity (i.e. its material building blocks such as organs and tissues that belong to her own body) bear high importance in regards of ethics and systems theory in general. Damage to the physical body (the 'things') may bring the destruction of Perunika as a *particular* entity.

Thus, both the *systemhood* and the *thinghood* properties work in conjunction to describe the particular system (entity) we are modeling in a particular scenario.

3.1.2.3 Setness

The *setness* property here is important from a methodological standpoint. As we saw previously in 3.1.1.1 Positive definition of system, a system is a *set* which incorporates two *sets*: a *set* of things and a *set* of relations. Setness is a property that enables collecting certain particulars of the world (the systemsphere) by using particular criteria and thus creating a set.

Hence, when we use the criterion of primitives (things), we end up with a set of things of that system 'filtered' by that criterion. When we use the criterion of relations, we end up with a set of relations. Setness is the property that enables creating sets out of everything and anything in the universe, and is the basis of set theory. It is the quality or state of being a set.

Before in 2.1.6 Method of Levels of Abstraction I discussed the method of LoAs. A LoA is a collection (a set) of observables about a system. It is obvious that they are selected according to certain criteria (i.e. relation to a system, properties, etc.). Moral reasoning, since it is done by using LoAs, also includes making sets of 'things' in the environment (or the systemsphere) over which it can be applied.

3.1.3 Emergence

The systemic phenomenon of emergence is a subject of a long debate. In essence, what emergence is taken to mean is the spontaneous or willed creation of a pattern, shape, or a function, where these were not previously imaginable or explainable from (i.e. they are practically or actually irreducible to) the properties of the constitute elements of complex systems, or from their interaction patterns (Pereira et al., 2016; sect. 1.3. Emergence; p. 3) (Bedau and Humphreys, 2008; in (Valentinov, Hielscher & Pies, 2016)). Emergence is typically brought when discussing the famous, if somewhat mystical, expression: *the whole is more than the sum of parts* (Bertalanffy, 1969; p. 55).

Or, even more clearly, emergence "… refers to the appearance of higher levels of system properties and behaviour that even if obviously originated in the collective dynamics of system's components – are neither found in nor directly deductible from the lower level properties of this system. Emergent properties are properties of the 'whole' not possessed by any of the individual parts making up this whole" (Aziz-Alaoui & Bertelle, 2007; Preface).

It is a term used to describe a "vast category of spontaneous, and weakly predictable, order-generating processes" (Pereira et al., 2016; p. 3). Pereira and Saptawijaya go on to state:

> What does emerge? The answer is not something defined physically but rather something like a shape, pattern, or function. The concept of emergence is applicable to phenomena in which the relational properties predominate over the properties of the compositional elements in the determination of the ensemble's characteristics. Emergence processes are due to starting configurations and interaction topologies, not intrinsic to the components themselves [2]. This functionalism is, almost by definition, anti substance-essence, anti vital-principle, anti monopoly of qualia. (Pereira et al., 2016; p. 3)

In simpler words, though, we can take emergence to mean the arising of a functionality, shape, or a pattern out of the interaction of *things*, when these are set in a particular order and / or particular relations are established between them at a certain point in time.

Some common occurrences of emergence would be: extracting ('creating') information out of data; behavior and other processes of groups; mixing fundamental colors to get composites; ordering carbon atoms in different configurations so as to derive diamonds, graphene, biological systems and the heart of the pencil; appearance of a whole functioning human out of a single zygote; connecting four legs and a tabletop to create a table; synchronization and integration activity of neuronal ensembles to derive consciousness (see 2.1.3 Integrated information above); answers given by deep neural networks; connecting 4 wheels, an engine, a body and other parts to create a car; integrating the sentences of the current paragraph you are reading into a concept of emergence in your mind; and many more.

I have discussed the method of Levels of Abstraction (LoAs) in 2.1.6 Method of Levels of Abstraction above. Basically, emergent behavior is behavior that "arises in the move from one LoA to a finer level" (Floridi, 2011; p. 63). Some behavior cannot be understood or even perceived when looking at a particular LoA; but when moving to another LoA, they become apparent.

For example, the behavior of a particular muscle fiber cannot be perceived or understood when looking at the level of the whole hand doing something (as the movement of the whole hand might be a Cartesian product of all its elements, not easily granulated). However, if we focus on the muscle itself, and then go even further in detail to the fiber itself, we will understand what it does at that particular moment. In parallel, we will lose the more holistic picture of the hand since we typically cannot perceive or understand the movement of the whole hand by understanding the action of only a single fiber (we are missing the other elements to derive the Cartesian product).

Basically, emergence is the process through which a system is created out of a simple set of things. This also applies to the emergence of moral systems and moral norms ((Saptawijaya, 2015; p. 134) also, see below).

Strong and weak emergence

Emergence is sometimes taken to have two kinds: strong and weak. Spät cites C. D. Broad on what he considers strong emergence. According to Broad, we talk of strong emergence

> "if every aggregate of order B is composed of aggregates of order A, and if it has certain properties which no aggregate of order A possesses and which cannot be deduced from the A-properties and the structure of the B-complex by any law of composition which has manifested itself at lower levels" (Broad in Spät (2009; p. 161).

However, some authors deny that strong emergence can exist. For example, Spät also cites Strawson who denies that something that emerged cannot be traced back in a procedural manner. He claims that

> "If it really is true that Y is emergent from X then it must be the case that Y is in some sense wholly dependent on X and X alone, so that all features of Y trace intelligibly back to X (where 'intelligible' is a metaphysical rather than an epistemic notion). *Emergence can't be brute* [. . .] in the sense of there being absolutely no reason in the nature of things why the emerging thing is as it is (so that it is

> unintelligible even to God). For any feature Y of anything that is correctly considered to be emergent from X, there must be something about X and X alone in virtue of which Y emerges, and which is sufficient for Y". (Strawson in Spät (2009; p. 161)).

There is a general criticism that emergence is only in the mind and does not actually take place, chiefly because we don't know certain information about a model that can explain the behaviors that seem 'emergent' (see (Bedau, 2008). However, Bedau (2008) disagrees with this claim. He explains (weak) emergence as *explanatory incompressibility*, which is taken to mean the inability to 'compress' explanation of a property or a phenomenon by simpler ('short-cut') symbolics, terms and arguments. This is somewhat compatible with the general claim for emergence I already mentioned in the previous subsection, like irreducibility to the properties of the constitutive elements, but in a 'weak' sense. Weak emergence, according to him "*bars* in principle irreducible downwards causation" [emphasis original] (Bedau, 2008).

Therefore, Bedau goes on to define weak emergence as

> "If P is a macro-property of some system S, then P is weakly emergent if and only if P is generatively explainable from all of S's prior micro-facts but only in an incompressible way. This definition defines weak emergent macro-phenomena by the distinctive way in which we explain how they are generated from underlying micro-states" (Bedau, 2008).

Hans Jonas also has something to say against strong emergence. In his own words, "What looks like a leap is in reality a continuation; the fruit is presaged in the root" (Jonas, 1979 in Spät (2009; p. 169)). In general, proponents of weak emergence argue that since everything in the universe has to follow the same physical, systemic and informational laws—and if these laws are taken as deterministic in principle—then strong emergence (appearance of something our of nothing) makes no sense.

Emergence also happens when dealing with organizations and their behavior. Their existence and actions in the world are practically or actually irreducible to those of its components. Hence, Chopra and White:

> Business corporations, a species of artificial persons, may also be coherently described as subjects of the intentional stance. The corporation may be identified as an intentional agent by virtue of its corporate internal decision structure; this licenses the predication of corporate intentionality even though the internal decision structure incorporates acts of biological persons (French 1984, 44ff.). Indeed, a corporation's actions are often not amenable to a facile reduction to actions taken by its human "components." When a corporate act is consistent with established corporate policy (as in "Exxon bought the oil field because it believed that would increase profits, and its policy is to make such purchases where the projected rate of return on the investment exceeds 15 percent per annum"), it is describable as done for corporate reasons (as caused by a corporate belief coupled with a corporate desire), and thus as a result of corporate intentionality (French 1984, 44ff.). (Chopra & White, 2011; p. 16).

For the purposes of this work the difference between strong and weak emergence is not crucial. Whether emergence is practical (because of epistemological limitations) or actual, the fact remains that it does occur. Collective (complex) systems emerge out of the (inter)actions of their components, and they can be treated as either unified entities or as simple aggregates, depending on circumstance.

3.1.3.1 Emergence in ethics and morality

Similarly, in ethics the so-called *open question argument* aims to establish that ethical claims and questions have an irreducible significance on their own; a significance not derived from something else (Baldwin, 2010; p. 286). G. E. Moore tried to establish this by claiming that we simply know and recognize that good is good in itself; that it is not good because of something else. Even though we saw that what is good can be derived from value theory (see 3.1 Value theory in Chapter II. Literature review), this argument is semi-true in regards of emergence. Ethics, morality, moral systems, ethical claims, values, questions... these are all emergent properties and phenomena from the underlying substrate. They cannot be reduced to the interactions in the substrate without losing their distinctive properties. One cannot find morality when looking at the physical aspects of persons in a fight, for example. In this sense, they do have an 'irreducible significance on their own'.

Emergence also represents a certain system theory's 'challenge' for ethics (Valentinov et al., 2016). Although the authors are focused on corporate social responsibility, some of their commentary applies to ethics and specifically AI ethics. For example, corporations and other organizations and social systems are 'artificial' creations that emerge out of the interactions of their components (people and other). The authors mention Luhmann, who commented on social systems being *operationally closed* (Luhmann, 2013; p. 63). They are created for a particular purpose and only deal with things and information that pertains them. This position makes them complexity reducers, since, for instance, the Ministry of External Affairs does not have to deal or be concerned with the ecological or economic affairs of the state (except in a contingent manner).

However, this complexity reduction is a source of deep and structural moral issues. Since social systems and institutions are operationally closed, they often externalize expenses and contingent issues on 'others'. In short —they are not interested in what happens around that does not nominally concern them. At times this causes deep disturbances in various other systems (e.g. the ecology, resources, etc.) because they continue pursuing their goals (so-called *goal rationality*) regardless of the wider effects on society or the environment this pursuit instills.

Similarly can be said for AI entities and systems. They typically are endowed with goal rationality and operational closure; which means that they pursue their goals regardless of wider effects (or, in the language of *ethics of systems*, they are explicit on their *Achievement of Personal Goals* imperative, and implicit on their *Conservation of Personal Continuum*; see 4.3.2 The moral imperatives below). Arguably, as with organizations and enterprises, they do not account for some adverse effects extending out from their work.

So, a system commanding the metro transport in Barcelona might be programmed and be under the demand to perform increasingly better on punctuality; but in order to achieve this it would have to spend more electricity and even run over people on tracks to be 'goal rational'. More electricity might result in more coal burning, which results in more pollution, and more people, animals and plants dying from it. A simple 'goal rationality' in this case would cause wider, morally contentious issues that might even span a continent!

There are two approaches commonly offered to solve issues like this. One is informationally conservative, the other its opposite—informationally progressive. In the first case, the so-called *moral veil of ignorance* is suggested to deal with things such as these (see the discussion on 2.3.1 Substantial (ethical) implications in Chapter V. Discussion). Since AI systems will inevitably and always be cognitively (i.e. computationally) limited (see 3.4.2 Complexity), instead of trying to increase the overhead computational costs by including more observables they need to account for, we should instead attempt to remove observables that are not immediately and/or directly important. This will render the systems extremely goal-focused; and outsource (externalize) other concerns on other systems and human organizations that will be tasked to deal with them, and control and modify the AI systems accordingly to avoid wider implications.

The informationally progressive approach, in contrast, advocates for increasing the observables that systems and complexes (e.g. organizations) ought to account for. This is in order for them to include things and states of matters that they might effect, so as to attempt to inherently and systemically avoid adverse moral or other effects, and possibly cause positive ones. This seems to be the position of the authors (Valentinov et al., 2016).

However, this necessary needs to be in a limited manner, if possible at all (because of the previously-mentioned issue of complexity and computational limitedness).

Which one is better is a question of context. Some problems require being informationally conservative, others its opposite. For example, issues such as bias avoidance when using recommendation systems for criminal penalties would most likely require *less* information about the person—not more. Autonomous vehicles and trolley-type scenarios seem to me as suitable to be approached in an informationally conservative manner. Similarly for data and privacy protection. However, medical diagnosis and treatment would typically require *more* information in order to achieve better results. And some social policies might balance the two.

In any case, emergent properties in systems can and do have ethical and moral effects that present challenges which need be taken in consideration for any serious moral theory.

3.1.4 The systemsphere

The *systemsphere* is a term that marks a concept very similar to Floridi's *infosphere* (see 2.1.2 The infosphere above). In Floridi's *ethics of information*, the infosphere is 'everything'. Every informational entity is part of it, and it can be subject of study through the informational LoA. Similarly, the systemsphere contains every system (entity) that exists, have existed, and will exist. It is a kind of set whose elements are all systems.

The question that can be asked here is: why is there need to discuss the systemsphere here at all. The systemsphere can have some purpose in discussing AI ethics, and ethics in general. For example, if we go back to IE's OOP model of moral action (see 2.2.4 Moral action above) we can see that both the agent and the patient are enclosed in a personal informational shell, which contains information and data they hold about the world. However, since they are not omniscient, they have a limited (and necessarily biased) view on the world. This is because there is a lot outside their shell. Everything else, for them and for us, can be treated as part of the wider systemsphere.

In the previous subsection I discussed about the challenge systems theory presents to ethics, mainly because 'artificial' systems and entities are purposefully unaware of the wider effects of their processes. Thus, when doing analysis, we can treat the wider universe as part of the systemsphere which exists and can be affected by actions of particular systems, even without being aware of it (see the example of the metro-commanding system above). In this respect, besides flourishing of individual systems, the wider moral dimension also includes flourishing of the whole systemsphere as a system itself.

In short, the systemsphere contains everything that is a system; and at times can be conceptualized as 'something' (i.e. a system that integrates all other systems) to account for processes and effects with moral effects that pertain to it. Being a system itself, it has the right to flourish as well.

3.2 Constructivist realism

Constructivism is the view that 'reality' is constructed by us, cognitive entities, typically through experience and/or analysis. This construction is based in making more or less arbitrary separations in an otherwise continuous and indistinct universe, and does not (or at least cannot be confirmed to) reflect actual state of matters. Reality is either purely subjective, or intersubjective. In contrast, realism is the view that reality actually exists 'out there', in an objective manner and outside the observer. The only problem we have in this sense is our epistemological limitations, which bar us from figuring out the ultimate state of matters—but this is possible *in principle*. Improvements in measuring and other tools (i.e. methodology, modeling) provide us with better and better image of the objective reality of phenomena in the world.

In systems science, these two views translate into the following. Systems constructivism holds that systems are creation of our minds, and do not actually exist in the world. There are no distinctions in the world — distinctions are solely cognitively (for now: human) made. Systems realism, on the other hand, claims that systems do actually exist, and that we are discovering them in reality through our tools and methods. Distinctions (e.g. separate entities, systems or complexes) in the world do actually exist.

According to George Klir, most of the writings and authors in systems science 'undoubtedly' espouse the systems-constructivistic view (Klir, 2001; p. 23), at least when his book was published (2001; although he cites no research on the matter and seems to provide solely his personal impression).

My personal view tries to reconcile both views and leans towards the middle of it. It seems that part of the conflict is linguistic and conceptual. For example, what Klir seems to take as *systems* is actually their *models*—this is a linguistic issue. Additionally, I often use systems and entities interchangeably, because I consider them ontologically equivalent—an entity is a system, and a system is an entity. However, the common notion of an entity (as a physically distinct object in the world) is not always used in the same context as system, since systems are considered as complexes of (physically or otherwise) distinct entities.

For example, what we commonly have in mind when referring to a computer system is a set of physical entities connected in a new whole that has a certain new (emergent) way of functioning. But that system can in parallel be considered a new entity, even in an ontological sense and not simply epistemologically. The reason is that if the system has novel (emergent) properties and ways of affecting the environment and being affected by it, it is a new entity. For instance, we can treat a human being as both a unified whole, and as a complex of organs and tissues, that are complexes of molecules, that are complexes of atoms ... Being made of complexes of atoms does not deny that a new, emergent system does really exist as a human body and being. To do so would be extremely dehumanizing (morally abhorrent), but also incorrect. That being said, to treat the human being as solely a unified whole is far from a correct and scientifically-supported view, and can also be dehumanizing (morally abhorrent) and morally biased (for example, by not taking groups and organizations in moral account). Additionally, even though gamma rays can pass through the body just like we can pass through the premises of a company or a territory of a state does not mean that both the body, the company or the state do not exist as separate entities.

It seems to me true that we can make constructions in relation to systems, but these are not the systems (or, if one prefers, the entities) themselves—only *models*. We attempt to create models that are in a homomorphic relation to the actual systems (i.e. to entities, complexes, scenarios, or else), but don't always succeed. We can also construct purely imaginary creations (like we do in fictional literature, for example) that may not even purport to represent reality (at least consciously and explicitly). These last are purely imaginary creations that we construct *seemingly* 'out of nothing'.

However, I do not agree with the claim that distinctions (and thus systems; see 3.1.1.2 Negative definition of system above) do not exist in the world and are purely human constructions. Red is distinct from blue, these shoes are distinct from my feet, and my body does not magically melt into the wall when I touch it. The constructivist claim is that there are no distinct systems in the world (like people, buildings, rivers, groups, states, etc.), and everything is a single, seamless and continuous whole. This is obviously incorrect. Distinct things exist in the world, and still they are *at the same time* integrated into the grand unified whole (the universe). This is why Luhmann (2013; p. 63) states that "The distinction between system and environment is produced by the system itself. This does not exclude the possibility that a different observer observes this distinction, which is to say, observes that a system exists in an environment".

Yet we might be pragmatically limited in our capacity to gather data and extract information about these systems anyway, and hence be forced to fill out gaps in our models with suitable imaginary constructs and complexity-reducers (e.g. assumed median values). We are most probably often also limited by being included in a greater complex with the system we have under analysis, while not being aware of this. The world is most likely not *experiencer-independent*, as the observer effect in physics and quantum mechanics demonstrates (e.g. in the double-slit experiment; see also (Von Foerster, 2003; p. 288)). This can weaken our capacity to discover the 'true' state of matters in an objective manner, in contrast to what the realists claim.

From the outside, a system *is the difference* between the system and its environment (see 3.1.1.2 Negative definition of system above). From the inside, however, the system has a certain particular *structure* made out of the set of things and the set of their relations. This structure defines the system internally (on the internal side of the boundary). And it is by this structure that we can see how that system is different from its environment (which does not belong in it). Jean Piaget is often times mentioned as a constructivist (Klir, 2001; p. 21, footnote), but even he has the following to say regarding structure:

> The discovery of structure may, either immediately or at a much later stage, give rise to formalization. Such formalization is, however, always the creature of the theoretician, whereas structure itself exists apart from him. (Piaget, 1970; p. 5)

This 'formalization' that Piaget talks about is the creation of a model of a system, but the structure (the system itself) does exist regardless if we do or don't create a model of it.

To conclude, a constructivist realism (or realistic constructivism) approach seems to be the best way forward (see also (Cupchik, 2001)).

3.3 Systemic Being

I have discussed previously in 2.1.2.1 Informational entities and 3.1.1 A system about the existence of (informational, systemic) entities and systems. In short, what exists is *something*, as opposed to *nothing*.

A something that exists is delimited in space and time[23]. I already discussed in 2.1.5 Cognition and reasoning that in order for *something* in logic to be meaningful, it has to be different from everything else—*a* is *a* because everything else is *not a*. Difference is essential for reasoning, and as we will see below, also for existence[24]. However, this delimitation (the difference between an entity and its environment) is not always clear-cut, both from an epistemological and an ontological perspective.

Epistemologically, we might not be aware where exactly, and *if*, a system ends and its environment begins. Ontologically, a system might be different from another on one level of abstraction (i.e. different people or AI entities on a street; or employees of a company), but be conjoined on another (i.e. the aforementioned being participants in the same moral scenario; or the employees working on the same company project).

Virtually all systems are semi-closed/open, which means that they are in a metabolic relationship with their environment (Valentinov et al., 2016; p. 600). Some systems might seem 'closed', but this is just a mirage since everything in the universe is interconnected. However, for pragmatic reasons it is good practice to limit the exploration of connectedness between systems, not least because it is simply computationally impossible.

[23] This of course does not apply for the totality of existence i.e. the whole universe (the systemsphere) who is by definition unlimited.

[24] It pays to be reminded of Integrated Information Theory's claim that a mechanism *exists* from the intrinsic perspective of a system only if it plays an irreducible causal role—it is a difference that makes a difference (Oizumi et al., 2014).

The boundary of a system

How are we to determine a ‘pragmatic’ point where the boundary between a system and its environment is? We can do this through several approaches.

The first is *differential*, by reminding ourselves of Luhmann’s definition of a system: “a system *is* the difference between system and environment” [emphasis original] (Luhmann, 2013; p. 44). A system ends where it *seemingly* stops being affected by actions of other systems; and for many practical purposes this is enough. Additionally, “The distinction between system and environment is produced by the system itself” (Luhmann, 2013; p. 63). The second is by excluding certain facts (i.e. observables) about a system that seem ‘irrelevant’ or at least not immediately relevant (this is reducing complexity based on descriptive information; similarly to the previous method) so that the boundary can be set somewhere. The third is contrasted with the previous one because it works by adding new variables that reduce uncertainty-based complexity (e.g. additional measurements of a boundary). The fourth and final is by making statistical assumptions (i.e. averages) of where the boundary is. All the previous methods are defeasible[25] and open to modify derived conclusions when new data is available. However, as Klir points out, systems are often resistant to simplification strategies, even though simplification is inevitable (Klir, 2001; p. 161).

Regardless of which strategy we use to determine the boundary of a system, we will arrive at a certain point (and by extension a line, a plane, etc.) in space[26] and time where the system ends and its environment begins. This boundary determines where the existence of a system is located—its Being. We can represent this boundary symbolically, in example, by drawing a closed line around a system in a graphical representation.

What is Being

And within this Boundary, Being is uniform (Floridi, 2013; p. 65) and integrated (see 2.1.3 Integrated information) into a coherent whole. Here uniform, in contrast to Floridi’s uniformity, means that Being cannot contain a total contradiction in itself that manifests at the same time and place. To be integrated means that it is more than the simple sum of it parts, which means that it has certain emergent properties. This is a very important conclusion for the further work here.

The Being of a system is its existence, in a qualitative and quantitative sense. It contains certain properties that characterize it. These are **structure** (comprised of the set of things and set of their relations), **cause-effect constellations** (that include, for example, systemic goals; but also any kind of implicit functioning of a system), **boundary**, and **position in space and time**. They are all intimately connected.

Being has two perspectives: **internal** and **external**. The internal perspective is subjective, and it comes from panpsychistic properties (see 2.1.4 Panpsychism and also 2.1.3 Integrated information above); while the external can either be objective or intersubjective (depending on the theory; see 3.2 Constructivist realism above). Every Being has them both, regardless if the system itself or external observers are aware of it. Hence, algorithms, machines, people, organizations, trees, mountains, planets and all other entities have a quality of internality (what it is to be that entity from the inside), and a quality of externality (what that entity is from the outside, as difference from its environment that makes a difference, and for other entities).

Furthermore, if the Being is aware of itself (becomes self-conscious), it can acquire a sense of selfhood (see below).

[25] The reason for this is that virtually all methods we use are heuristical *in practice*, owing to cognitive and/or computational limits; see 2.1.6.1 Is the LoA method essentially a heuristic? and 1.1.2 Is EoS too reliant on heuristics?

[26] We should be careful not to take the simplistic, physical meaning of space as relevant here. Space can be both physical, but also informational, systemic, conceptual, mental… depending on our level of abstraction.

Selfhood

Some Beings have a sense of self. Obvious examples are people, but probably also some animal species. Potentially, AI entities might get to be designed with, or develop, a sense of self (Chrisley, 2008).

The sense of self does not seem to be a unique cognitive process (or aggregation of processes) in terms of *quality*, but in terms of *subject*. In a system (e.g. a brain) cognitive processes can have a certain focal point which is their subject. An animal can be focused on a prey or a predator, for example. However, if those mental processes somehow turn inwards (as can be induced to in the cases of suffering and pain[27]) and make other internal processes or even themselves as a focal point, a new type of consciousness arises—self-consciousness.

To finish with this section, I will cite Freya Mathews and her commentary on selfhood which is closely related to systemic Being:

> "This geometrodynamic plenum is holistically rather than aggregatively structured, and those internal differentia which are not only stable in their configuration, but actively self-realizing, qualify as what I call selves. Selves are defined, in systems-theoretic terms, as systems with a very special kind of goal, namely their own self-maintenance and self-perpetuation. On the strength of their dedication to this goal, such self-realizing systems may be attributed with a drive or impulse describable as their conatus, where conatus is understood in Spinoza's sense as that "endeavour, wherewith everything endeavours to persist in its own being". (Spinoza 1951, Part III, Prop VI, Proof).
>
> Selves then enjoy a real though relative individuality even though they exist in the context of an undivided whole. Since they proactively seek from their environment the resources they need to actualize and maintain their structure while at the same time resisting causal inroads into their integrity, they count, ontologically, as individuals, even though they are not separate substances, but disturbances within a global substance. Moreover, the interference patterns which create these relatively stable configurations in the plenum are relational: it requires a very special "geometry" in the surrounding field to create the conditions for such self-perpetuating "vortices". The paradigmatic instances of selfhood, in the present sense, are of course organisms, constituted in the relational matrices of ecosystems. The systems-theoretic criteria of selfhood – self-regulation, homeostasis, goal-directedness and equifinality – may also turn out to apply to higher order biological systems, such as ecosystems and the biosphere. Indeed, it may be argued that the cosmos itself satisfies these criteria, since it is necessarily self-actualizing and self-regulating, and its self-structuring follows the relational dynamics of systems. (The details of this argument can be found in Mathews 1991.)" (Mathews, 2011; p. 5)

It can be safely taken that selves are important for ethics and morality. They are typically (but not necessarily) the result of more advanced cognitive entities, and they provide internal awareness of moral phenomena such as suffering, pain, empathy, intersubjectivity, rule discovery/establishment, and similar.

3.3.1 Structure, and wholeness

A Being is an integrated and uniform systemic whole. As mentioned above, internally Beings have structure which is determined (primarily) by the specific constellation of the set of things and set of their relations. Of course, cause-effect constellations also arise out of the structure since specific inputs tend to cause specific outputs[28] exactly because of the specific structure of the system (see 3.3.1.1 Structure is (causal) constraint

[27] This is why suffering has formative effect on the ego, since it helps the cognitive entity discover the boundary between the self and the outside world.

[28] We should be careful here to avoid conflating epistemological with ontological issues. If we have perfect resolution of a system we would be capable of determining what exact output will particular inputs give (if any). However, this is computationally/cognitively unfeasible, as will be discussed below in 3.4.2 Complexity. This is an epistemological limit, not an ontological fiction; although a limit that cannot be overcome because of how quantum mechanics works (Harshman, 2016; p. 11).

below). This structure is a unique pattern that comprises the uniqueness of every Being (see 3.3.2 Being as pattern below).

Elements that are in the set of things have particularities that determine the structure. For example, a military unit can be composed out of several soldiers having different positions in space and time, and specialities e.g. different weapons and functions. The relations also determine the structure. The soldiers in the same military unit have different ranks that determine things such as where they will be positioned, how they will behave, and what kind of power they will exert over whom.

For military purposes the upper echelons (i.e. generals, admirals and the like) will not be dealing with the particular members of hierarchically very low units (i.e. squads) and what they do, unless there is a very good reason to do so. They would look at them in a simplified manner and regard them as cohesive entities with a particular position in space and time that can perform an array of functions. Similarly can be said for individuals (complexes of organs and tissues), families (complexes of individuals), organizations and similar.

Structure also typically has certain resistance to external pressure (inertia). Buildings, planets, mountains, bodies, knives, and organizations all can withstand limited external pressures upon their structure before they stop functioning as they did previously. This capacity is not limitless, as we are very well aware (see 3.3.3 Injury and destruction of Being below).

Additionally, too much rigidity can result in less internal flexibility of a system. This might mean that a system is not capable of adjusting to environmental or internal change, and can lead to unwelcome results e.g. its destruction. It can also mean that too rigid of a system is not usable for applications where flexibility is required i.e. we cannot use a stone when we need a bicycle tire.

Too small resistance to external pressure (rigidity), on the other hand, might mean too high volatility. Such a system might break down or abruptly change in the way it functions when a relatively weak external pressure instills a change in its structure. An example would be a drop of water entering a computer and breaking it down because it hit a particularly sensitive and important component.

Where exactly is the ‘golden point’ between rigidity and flexibility depends on context. Some systems need to be highly internally volatile and dynamic, but then we keep them shielded from many destructive influences e.g. computers and bodies. Others are very rigid and we don’t defend them as much, but we don’t typically use them for applications where flexibility is required e.g. concrete and stones. Some are rigid in certain parts (i.e. bones), and flexible in others (i.e. digestive organs).

Wholeness

Finally, structure gives rise to wholeness, where a certain entity becomes something more than the simple sum of its parts—a whole. From a particular point in time onward the parts become integrated in this whole, and we can freely speak about a new entity thereon. Hence, Piaget:

> “That wholeness is a defining mark of structures almost goes without saying, since all structuralists-—mathematicians, linguists, psychologists, or what have you—are at one in recognizing as fundamental the contrast between structures and aggregates, the former being wholes, the latter composites formed of elements that are independent of the complexes into which they enter. To insist on this distinction is not to deny that structures have elements, but the elements of a structure are subordinated to laws, and it is in terms of these laws that the structure qua whole or system is defined. Moreover, the laws governing a structure’s composition are not reducible to cumulative one-by-one association of its elements: they confer on the whole as such over-all properties distinct from the properties of its elements” (Piaget, 1970; p. 6).

In order for something to be a *whole* it has to differ from its *parts*. This is how we discover that there is a whole after the fact that it came into being. I already mentioned before that difference is important, not just for reasoning, but also for existence itself. The whole must be making a difference above and beyond what its parts are (capable of) making. Thus, Oizumi, Albantakis and Tononi:

> "Recall that IIT's information postulate is based on the intuition that, for something to exist, it must make a difference. By extension, something exists all the more, the more of a difference it makes. The integration postulate further requires that, for a whole to exist, it must make a difference above and beyond its partition, i.e. it must be irreducible" (Oizumi et al., 2014; p. 10).

These are the emergent properties of wholeness.

This, of course, does not imply that the whole is doing a greater difference than its parts *at all times*. Sometimes, at particular moments, its parts can do greater difference. This also does not mean that the whole suddenly ceases to exist. In a sense, sometimes the whole can be 'paused' (or not considered) for awhile.

Striving for wholeness, and the calculus of Being

Systems often have aims (goals), whereby aims are state(s) of the internal and external world that they want to see come about (for a more detailed discussion see 3.4.3 Goals below). Therefore systems are, by definition, not complete when their goals are not achieved. That is to say, they are not (yet) a complete whole, internally. Logically, when a system achieves all its goals in the world, it finally becomes a complete whole—internally.

Since we will see below that goals are (ranges of) values of observables that a system wants to see come true, and since observables can be written down as either Boolean discrete values (true/1 or false/0) or as closed intervals between them [0, 1], by extension the wholeness, or completeness, of a Being can also be written down as closed interval between 0 and 1 → [0, 1].

For example, if a Being has achieved all its goals it has become a complete whole with the value of 1. A Being that has achieved half of its goals has a value of 0.5. A Being that has achieved almost nothing has a value approaching zero, and a Being that did not achieve anything has a value of exactly 0. The amount of difference a goal makes on a Being can be obtained both in an objective/intersubjective manner (e.g. statistically); and in a subjective manner.

We will see in 4. Towards Ethics of Systems below that the above will help us form an ethical calculus that AI entities can perform to carry through sound moral decision-making.

3.3.1.1 Structure is (causal) constraint

In the previous section I mentioned that cause-effect constellations arise out of the systemic structure. The reason for this is that internally systems follow certain laws that govern transformations which happen within them. The connection between an input and an output is an internal transformation. Thus, Piaget:

> "As a first approximation, we may say that a structure is a system of transformations. Inasmuch as it is a system and not a mere collection of elements and their properties, these transformations involve laws: the structure is preserved or enriched by the interplay of its transformation laws, which never yield results external to the system nor employ elements that are external to it. In short, the notion of structure is comprised of three key ideas: the idea of wholeness, the idea of transformation, and the idea of self-regulation" (Piaget, 1970; p. 5).

Piaget also continues further down the book by saying that it is also a delimitation of "*possible* states and transformations" [italics original] for a system (Piaget, 1970; p. 38). Similarly, Fultot (2016) comments that "the very state of the system [...] counts as a constraint".

Therefore, structure is a causal constraint. This means that it tends to limit the repertoire of links between inputs and outputs—the repertoire of what can happen inside a system—to a certain, theoretically-limited set. Because of this, this constraint enables the *particularity* and *repeatability* of the pattern that represents a particular Being (entity, system; see also 3.3.2 Being as pattern below).

For example, when I use a calculator, a car or a gun, I always expect them to act in a certain manner within a limited repertoire. Whenever I input *2 + 2* = in a calculator I expect to see 4 as a result, not a drawing of Picasso. My car and gun, however, cannot normally be used as a calculator (that is, are constrained in this sense). Therefore, the calculator has a certain particularity in internal structure different from those of the car and the gun. But if I have the scheme, I can build virtually identical calculator, thus copying (repeating) its internal structure and pattern (especially its systemhood; see 3.1.2.1 Systemhood above); just as well as when I copy a file containing a document or a song.

That structure is a causal constraint is also reflected in Integrated Information Theory:

> "… a mechanism in a state generates information only if it constrains the states of a system that can be its possible causes and effects – its cause-effect repertoire. The more selective the possible causes and effects, the higher the cause-effect information *cei* specified by the mechanism" (Oizumi et al., 2014; p. 3).

If you remember, IIT includes the notion of MICE—maximally irreducible cause-effect repertoire. For IIT, a *mechanism* is anything that has a *causal* role within a system. Thus, a mechanism can be a neuron in the brain, a logic gate in a computer, a lever that opens a door, an electromagnetic switch… The focus here is on its causal role.

> "In IIT, information is meant to capture the "differences that make a difference" from the perspective of the system itself – and is therefore both causal and intrinsic. These and other features distinguish this ''intrinsic'' notion of information from the "extrinsic", Shannon notion [...]. Information as "differences that make a difference" to a system from its intrinsic perspective can be quantified by considering how a mechanism in its current state s_0 constrains the system's potential past and future states" (Oizumi et al., 2014; p. 6).

Causal constraints create patterns, and Beings as patterns are the subject of discussion in the following section.

3.3.2 Being as pattern

When trying to consolidate a definition of 'pattern', the following synonyms appear most often: *model*, *arrangement*, *order* and *example* (Dictionary.com, 2019a). We can disregard 'example' for now (since it refers to the past) and stick to concepts of the present.

More technically, a pattern is a particular (in principle and theoretically) repeatable sequence of changes. In the most basic notion these changes, and thus the pattern, may be represented in a totally contrasted, simple, binary form i.e. 1s and 0s (see also 2.1.5.1 Reasoning as the perception and manipulation of differences).

As we have seen above, Being has structure, cause-effect constellations, boundary, and position in space and time. They give the uniqueness of each particular Being. All these can be described as being themselves, or being in a relation to, the three synonyms—model, arrangement, order. Therefore, Being can be regarded as a unique *pattern*.

That Being is a pattern is not a radical claim. In many[29] philosophical, religious and mythological traditions the notion of 'soul' fits closely to the notion of a Being-as-pattern. Hence, the soul of Christ can inhabit a person; a person's soul can continue on even after the demise of the physical body; and Hinduism and Jainism hold that *every* living thing not merely has, but *is* the soul itself—the Atman. The pattern of a Being is an emergent derivative in the form of a "distinctive kind of complex, macro pattern in the mind-independent objective micro-causal structure that exists in nature" (Bedau, 2008).

Similarly, Jordan B. Peterson would say:

> "A spirit is a pattern of Being. Patterns can be transmitted across multiple substrates: vinyl, electronic impulses, air, vibrations in your ear, neurological patterns, dance—it's all a translation of what you can describe as spirit" Peterson (2017; time: 01:35:15).

Sometimes the notion of soul is equated with mind, in order to refer to consciousness. In this sense, Integrated Information Theory (see above) states that: "An experience (i.e. consciousness) is thus an intrinsic property of a complex of elements in a state: how they constrain – in a compositional manner – its space of possibilities, in the past and in the future" (Oizumi et al., 2014; p. 14).

Being-as-pattern can be seen in things such as ideas, works of art, communication, identity, objects, memory and storage, copying and copyright… For example, a song can be duplicated on (and deleted out of) a CD, a USB stick drive, a hard-drive in 'the cloud', from my phone, or from an LP—but it is still the same song because the recording and reproduction follow the same pattern, that (attempts to) create the same *pattern* of sound waves.

Being as pattern through space and time

This notion of Being-as-pattern fits the whole concept of dynamic, ever-changing, temporary existence. This is why the second moral imperative of systems is Conservation of Personal Continuum, which is based on retaining a *pattern* through time and space that ought not be cut, but instead ought to remain continuous (see 4.3.2.3 Conservation of Personal Continuum below). Therefore, Being is also a pattern through time and space.

This pattern is retained even when small changes that do not cause substantially different functioning of a system are introduced. For instance, a person's awareness goes away when in deep sleep because of breaking of neuronal integration, but that doesn't mean that suddenly that person is gone, never to be back. When the person's 'presence' comes back in his wakeful state, the person is ('re')integrated. The person feels being the same person, responds when called by the same name, and reports having memories of being awake before going to sleep.

This is not conclusive evidence that the person is the same person after (re)integration, however. It might be that, since there is a break in continuity (however brief), the old person (system) has ceased to exist and a new one was formed. This has some 'troubling' implications, though.

If a change in person's integration breaks the very person (system) itself, that would mean that indeed <u>any</u> change in systemic structure (e.g. a new element, old elements in a new configuration) will break the person (system) from one moment to another. This would, furthermore, imply that there is no such thing as a particular person (system), which would seem to support the Buddhist claim of 'Anātman': that there is actually no such thing as "I" (ego, me) and identity because of continuous change.

Yet, when a particular person loses a part of their brain, but continues to identify with their name and other personal data, we (and they) don't seem to claim that they are not the same person anymore (in contrast to

[29] But not all—for example, Buddhism holds that there is no such thing as soul. Life is a constant ever-changing flux, an impermanence that cannot accommodate soul (Atman), and thus there is no-soul (Anātman).

Buddhists and similar schools of thought). The same applies for systems that continue to work similarly to before even after a change: a person losing an arm or a leg, or gaining an artificial limb; a country losing or gaining part of its territory; an ISP losing part of its data servers, or adding new ones; an organization laying off a whole division of workers, or hiring new ones. In all these cases the person, the country, the ISP, and the organization continue to exist, albeit differently.

Thus, even though a system can change through time, if it continues to function similarly to before and to identify as the same system, it could be said that the system at each particular moment *both* is and is not the same system. A person today is not the same person it was 10 years ago—but in a sense, *it is* the same person since there is a string of continuum throughout all that time. This is the basis of the **Conservation of Personal Continuum** moral imperative, which is implicit or explicit in systems (see 4.3.2 The moral imperatives below).

This continuum is a causal constraint through time. It specifies that there has to be a significant, but not total, causal connection between a system's Being from the first moment of conception to now, or to a point in the past when the system stopped existing. If such a continuum exists for a particular system, we can say that it is (was) the *same* system over that time.

3.3.3 Injury and destruction of Being

A Being is something, and that something can be injured or destructed as we are well aware. Injury or destruction is typically a (morally) *negative* change for the system, unless they are expected in the particular context. They cause change in a system's structure or boundary, after which its capacity to achieve goals[30] will be somewhat or totally impaired.

Since the entity has a particular kind of pattern (order) the above type of changes introduce unwanted or unexpected disorder—either by modifying a system's things or their relations, and by extension, the cause-effect constellations.

This disorder is sometimes labeled 'entropy', although Floridi (2013; p. 65) is very vocal about this being an unfortunate linguistic choice that is hardly rectifiable now. Thus, it bears to point out that under consideration here is metaphysical entropy, not thermodynamic or information-theoretical.

Entropy has a particular effect on Being. Since we have seen that something (i.e. an entity) is only something (i.e. exists) if it is a difference from everything else, entropy 'acts' against that by erasing the differences between an entity and its environment. The more entropy is introduced, the more an entity is injured. Entropy is a movement towards non-Being (recall the discussion in 2.2.1.2 The Evil) This is a process that, if continued, can inevitably only end in destruction of an entity.

Entropy is, however, part of the life of any and all systems. Systems can resist entropy for a limited amount of time, but as we know very well, no system can survive it indefinitely—and all systems will eventually cease to exist (except maybe the systemsphere). Since (spontaneous) entropy is a constant reality, it brings about constant change in systems. If they are to survive for an extended amount of time, their Being has to maintain a dynamic order (in contrast to a static, rigid one). Moreover so, in order for any Being to be able to maintain personal negative entropy, it needs to 'export' positive entropy outside (by which local and global entropy increases). This is why there is a difference between destructive and constructive entropy (for further discussion see 4.2.2 The Bad and the Evil below).

Injury and destruction can be stochastically, unintentionally, or intentionally introduced to a system. This has very important moral implications for the *ethics of systems* theory that I formulate at the end of this chapter

[30] Of which the conservation of personal continuum can be one.

(see 4 Towards Ethics of Systems). Stochastic introduction of entropy can be taken as moral bad or tragedy. Unintentional introduction can be either moral bad, or evil in cases of severe negligence. Finally, intentional introduction generally makes moral evil (except in cases of self-defense[31]).

Since injury can be gradual, it can be in principle mathematically represented on a period between 1 and 0 → [1, 0] and then multiplied by -1 to make it a negative change. This is an effect on a system's existence and flourishing (what I call Quality of Life in 4 Towards Ethics of Systems below), either by negatively affecting its conservation of personal continuum, or its achievement of personal goals. The end result will be a negative difference (Δ). Thus, $x_{moral\ process} = (-)\Delta QoL$.

More formally, when a system is injured there has been instilled an undesired change in some of its observables. Their values have been modified to a state outside normal operational limits of that particular system (Ashby, 1999; p. 197), at least momentarily. Whether the system survives this change is dependent on its capacity to resist such changes at that moment.

3.4 Systemic deliberation

All systems deliberate (cognize), because all systems have at least rudimentary psychic capacity (see 2.1.4 Panpsychism above). What information they hold at any particular moment is the result of integration, and when their structure changes this integrated information also changes resulting in changes in their subjective experience (see 2.1.3 Integrated information).

Some have specialized components that deal with cognitive processes, while others perceive and think with significant portions of their whole Being. Systems also differ in the speed of cognition, their capacity to hold and process data, and all other parameters of cognition.

Cognition always happens in the present moment, but its subject can be either the present moment, the future, or the past. Entities can reason about what might happen or what did happen, and the stronger their cognitive capacity the more capable they are of reasoning about the future or the past (besides the present). The present, however, cannot be reasoned about on its own (i.e. without regard to the past or the future) because it is a *prima facie* occurrence that can only be experienced[32].

By extension, moral entities can also have the capacity to reason. When finding themselves in a moral scenario, they can become aware of the situation, consider the possible courses of (in)action, and decide which one to take (if any). Moral steermanship and moral communication is part of reasoning, particularly of cybernetics (see next section).

Cognitive capacity is one factor (and a multiplier at that) of power in a moral situation for an entity—the higher the cognitive capacity, the bigger potential for power the entity has (up to the maximum possible power the entity can have in the context). But the higher the power, the greater the moral responsibility. Of course, entities are inevitably limited in their cognitive capacity because of complexity, with an upper boundary—Bremermann's computational limit (see 3.4.2 Complexity below).

Systemic deliberation is a very important property for morality, as we will see below.

[31] Self-defense can take different forms. For example, classical self-defense is physical reaction to (a threat of) physical aggression. However, legal and moral enforcement of agreements can also be taken as self-defense, as well as, when societies arrest and put suspected and proven criminals in jail.

[32] Which is one of the primary teachings of Buddhism and Daoism.

3.4.1 Reasoning and cybernetics

Reasoning is also closely connected to cybernetics. Merriam-Webster defines *cybernetics* as “the science of communication and control theory that is concerned especially with the comparative study of automatic control systems (such as the nervous system and brain, and mechanical-electrical communication systems)” (Merriam-Webster, 2019a). Ashby (1999; p. 1) also mentions Wiener’s definition: ““the science of control and communication, in the animal and the machine”—in a word, as the art of *steermanship*” [italics original]. So, cybernetics is about communication and control in systems. These can be either performed by dedicated component(s), or by the whole system itself.

> “At some level organisms, machines, or other objects in the world are responsive not merely to the fundamental laws of physics, but also to information that they obtain from their environment that they interpret in some way and that may affect their behavior” (Nugent, 2018).

In general, systems that are sophisticated[33] about the achievement of their goals (and particularly those whose goals include self-preservation), especially in complex and dynamic environments, deal in cybernetics. Cybernetics is a subset (a special case) of reasoning. This equally applies regardless if a system has a dedicated communication and control component, or coordinates, regulates and controls with significant parts of its whole Being.

Examples of cybernetic components of systems are control and communication systems both in biological entities, e.g. brains, endocrine system; and in artificial entities and machines, e.g. autopilots *and* pilots, national assemblies.

As mentioned before, control and communication in moral scenarios is very important for morality. As Von Foerster (2003; p. 289) himself points out, when a cognitive entity (i.e. a brain) accounts for its own activity—a process that is in the focal point of ethics, particularly virtue ethics—this represents *second-order cybernetics*: cybernetics about cybernetics. Internal processes of morality i.e. building frameworks, making decisions with moral effects, considering those effects, avoiding harm and promoting good, self- and other-preservation, etc. all depend on introspective examination—a sort of cybernetics of cybernetics.

How is cybernetics important for this work

When speaking about AI entities and morality, we also can explore their cybernetics. They can either have explicit or implicit moral cybernetics. Furthermore, they can also have either a dedicated component that deals with moral cybernetics, or they can do this with significant part of their Being.

In Chapter I. Introduction I made the case that the widespread introduction of AI entities in societies and their steady movement towards increased complexity and capacity to act will bring about a plethora of novel moral issues. Where their employment has the possibility to cause significant moral effects they will need to take these issues into account, either explicitly (as an explicit factor in their programming) or implicitly (such issues to be considered by their creators, employers, users, etc.).

The matter of fact is that the goal of this thesis is exactly to contribute to the above.

3.4.2 Complexity

Complexity is a concept that probably arose tens of centuries ago, ever since people started noticing that entities and phenomena in the world are not always homogeneous but often times are composed of multiple,

[33] Sophisticated, as in, have the capacity to take into consideration a plethora of observables that might be in complex relations.

semi-independent components. Therefore, entities can be complex, and it bears to look at them in this way because their components can have separate presence, effects and behaviors.

The word *complex* is an adjective that refers to something which is “composed of many interconnected parts; compound; composite” (Merriam-Webster, 2019b). Being an adjective, it can be used for entities and phenomena of the world. Hence, we can speak of complex societies, relationships, organizations, chemicals, carbohydrates—or more generally, of systems. When speaking about complex systems we can also use simply: *a* complex; because systems are at the foundations of structured reality.

Complex systems are those whose components are, at least in part, other systems[34]. For example, we can analyze a building and discover the bricks, concrete, steel bars, insulation, electrical wires, plumbing etc. that make it. If we go in further analysis we will see that all these components are complex systems themselves, composed of other complex systems ... In the other direction, that building can be part of a neighborhood, which is part of a municipality, which is part of a city, which is part of a region, then a state, the planet, *etcetera*, ending with the universe itself (see the illustration in 3.1.2.1 Systemhood, Hierarchy of systems).

> “In some contexts, complexity is a desirable property, i.e., we search, within given constraints, for systems with a high degree of complexity. Cryptography and the design of random number generators are two typical examples of such contexts. In some situations, a certain degree of complexity is a necessary condition for obtaining some specific systems properties, usually referred to as emergent properties. Self-reproduction, learning, and evolution are examples of such properties. In other contexts, which seem to predominate in systems problem solving, we search for simple systems or attempt to simplify existing systems” (Klir, 2001; p. 159).

As we are well aware, there is a limit to our capacity of consideration (see below). A person can hold in working memory only a few mental objects, a computer has limited amount of memory and processing power per second, and an animal can track a limited set of objects during hunting. Similarly, we have a limit to how *complex* our models of systems can be. This translates in a limited capacity to consider observables of systems and track their states throughout time (remember the Method of Levels of Abstraction in sect. 2.1.6). The more observables (and their relations) we include, the more computational power we need to consider them.

This limitation is recognized by most systems and IT theoreticians, cyberneticians, computational linguists, theoreticians of mind, and anyone that has faced this unconquerable challenge. For example, the authors of IIT are keen to point out that

> “The present analysis is unfeasible for systems of more than a dozen elements or so. This is because, to calculate Φ^{Max} exhaustively, all possible partitions of every mechanism and of every system of mechanisms should be evaluated, which leads to a combinatorial explosion, not to mention that the analysis should be performed at every spatio-temporal grain. For these reasons, the primary aim of IIT 3.0 is simply to begin characterizing, in a self-consistent and explicit manner, the fundamental properties of consciousness and of the physical systems that can support it” (Oizumi et al., 2014; p. 24).

The final[35] theoretical limit on computational power is the Bremermann’s limit.

[34] Simple systems are those that are composed solely of primitives. However, whether these primitives actually exist (and thus whether *ontologically* simple systems exist) and can be discovered is an unresolved question of philosophy and physics.

[35] Although recent development in quantum computation might increase the upper threshold significantly. However, a limit necessarily must exist for all delimited entities in the universe (while not forgetting that the universe is already that total reasoner that has the maximum capacity which seems to be unlimited).

Bremermann’s computational limit

The Bremermann’s computational limit was introduced by Hans Bremermann in a 1962 paper (Bremermann & others, 1962), whereby he stated that “No data processing system, whether artificial or living, can process more than 2 x 10^{47} bits per second per gram of its mass”.

At first sight the number seems arbitrary. Why 2 x 10^{47}, and not 2 x 10^{50} or even 2 x 10^{1000}? Bremermann and the others in the team took into consideration Heisenberg’s uncertainty principle when trying to conceptually represent the recording and processing of different states by using discrete energy levels (whereby the size of precision of discrete levels is inherently limited). This is how the aforementioned number was derived.

He then combined this with the assumed age of Earth (10^{9}-10^{10} seconds) and assuming a computer with the same mass (less than 6 x 10^{27} grams) to arrive at the result of upper limit of 10^{93} bits processed for the period. For comparison, chess is estimated to have a totality of possible moves of about 10^{120}. Therefore, assumably, Earth throughout all its history could not calculate all possible chess moves of a simple 8 x 8 board and 6 types of pieces.

It is obvious that this limit can pose significant problems for anyone that attempts systems modeling. By estimations, even if we would like to select a single logic function of *n* variables that are under consideration, the maximum feasible number of *n* variables is somewhere around 310! (Klir, 2001; p. 147)

One solution is to simplify and disregard certain observables or even jump to conclusions, but this results in weakening the methods and turning them into heuristics (recall the discussion in 2.1.6.1 Is the LoA method essentially a heuristic?). This also applies to entities participating in or studying moral scenarios (see next subsection).

Hence, Harshman:

> “To embody the “intellect” (sometimes called Laplace’s demon) that knows and processes all information about the present state of the universe would, by some calculations, require more matter than the universe (Lloyd 2006). Further, chaos and complexity theory have shown that even simple systems can have exponential growth of uncertainty under dynamics, and generate long-range spatiotemporal correlations (Gleick 1987). Therefore even within classical mechanics deterministic prediction is always approximate. The finiteness of the observer guarantees the impossibility of total knowledge and the existence of ignorance, and this opens the door to probability (see Chapter 2) and randomness as useful concepts in physics” (Harshman, 2016; p. 8).

Complexity, simplification, heuristics, and bias

We have seen that there is an inherent limitation of cognitive capacity for delimited entities. The reason is that the reasoning or computational power that can be performed out per gram of matter is necessarily limited. However, this does not mean that they have to cease trying to pursue their goals in the world and that the battle is lost before it even started. Tigers still hunt, humans still build skyscrapers and drive cars, governments still project budgets, and trading algorithms are still employed for automated stock trading.

How can all the previous systems do these activities in a largely successful manner if complexity is such a pervasive feature of the universe? The answer is by using simplification strategies. Classical (mathematical) simplification that follows completeness is performed by statistical averaging and/or by abandoning factors in the calculation (e.g. disregarding observables). However, biological entities who don’t reason in explicit mathematics or logic tend to (additionally) use incomplete simplification techniques called heuristics.

A heuristic is a practical problem-solving method not guaranteed to be optimal, best, exhaustive or rational. Even though heuristics can derive erroneous or suboptimal solutions, at the same time, typically offer significant reduction in time and cognitive overhead.

For example, although remembering the multiplication combinations of the first million numbers would theoretically offer an advantage for mathematics, remembering them for the first 10 or 20 numbers helps solve most of the challenges of everyday life. Similarly, the genes of dogs bias them towards liking upright bipedal creatures, even though some of them don't like dogs and might hurt them. However, since the last happens significantly less often in comparison, that heuristic strategy was developed in their genes (but obviously not in the genes of tigers or lions who we hunted down mercilessly throughout history).

Except for completeness, there is no substantial difference between the two kinds of simplification strategies. They all abandon certain factors in calculation, or are biased for or against some, or jump to conclusions without an exhaustive calculation, in order to derive *good enough* solutions within the available time.

Hence, Simon:

> "The human species has survived and thrived in the world, simple or complex as it may be, not so much through the speed and power of its computational capacities, as by exploiting the fact that the systems of interest to it represent highly special cases that can often be analyzed by relatively simple means, provided their underlying structure is detected. This argues for a strategy of searching for that structure, of pattern induction-a skill that is rather highly developed in the animal kingdom-followed by special analysis and heuristic problem-solving search, rather than brute-force analysis of very general classes of highly interconnected complex systems" (Simon, 1977 in (Klir, 2001; p. 159)).

Weaver also comments on complexity and simplification strategies:

> "It follows from these observations that systems complexity is primarily studied for the purpose of developing sound methods by which systems that are incomprehensible or unmanageable can be simplified to an acceptable level of complexity. Such methods are crucial for dealing with phenomena of organized complexity [Weaver, 1948]" (Klir, 2001; p. 161).

We can see that model building is also simplification. When we regard a system simply as a whole without accounting for some or all of its components, we are performing simplification. Whether this is suited for the situation under study is a question of experience and best research practices.

In everyday life people, organizations, animals, robots and other systems also use another incomplete simplification techniques that we recognize under the label of *bias*. To be biased for or against means to have an irrational preference or repulsiveness for something—for a phenomenon or a (class of) entity of the world. Here 'irrational' means forming judgment without having a sound, complete and exhaustive argument.

Bias has rightfully earned its bad reputation, but has unrightfully been regarded *solely* as a bad cognitive approach. As we have seen before, complexity makes using heuristics (of which bias is a subset) inevitable. In addition, not all biases are bad. Following cultured biases towards cleanliness, well-mannered and measured behavior, or respecting traffic regulations all end up with better results *on average* than not.

We also have inherent, hardwired biases[36] that partake in our process of forming judgments, judgments which are not always directly connected to them. They have been evolutionary 'installed' in our genes through evolutionary pressures from the environment. Some of them are obviously maleficent or inefficient today, some of them are still as effective as they have ever been. Yet, when they were created they served a particular

[36] For instance, we have biases for things i.e. for sweet food, 0.7:1 waist-to-hip ratio in women, higher-than-average height in men, colorful shiny objects etc. We also have biases against things i.e. against foul smells, nonsymmetric bodies or darkness.

purpose to quickly and efficiently make judgments about certain phenomena in the world without entering analysis paralysis.

The problem with heuristics (such as bias or even abduction) is not that they can lead to suboptimal solutions. This is inescapable. The problem is when they are not updated in the light of new data, or when such data is actively avoided. Otherwise, heuristics provide a quick and efficient, if often suboptimal, methods to arrive at good enough solutions for particular situations. As such, they ought not be irrationally shunned away just because that *seems* like a good idea. Especially since all cognitive entities use them, even without being aware of it, and even when they deliberately attempt to go around them.

Similarly, in the field of practical ethics (see below) heuristics are also commonly utilized by moral entities, and we should embrace and work on improving them with purely rational methods, instead of hopelessly trying to replace them.

How is complexity important for this thesis

As I already mentioned before, systems have limited cognitive capacity because they are physically delimited. This also applies to AI entities. While they offer some advantages in respect of brute force computation and recently some pattern recognition, this does not mean that they are independent of bias and heuristics.

The matter of fact is, AI entities already employ heuristics on everyday basis. This is commonly done implicitly by the designers and employers of such systems, by feeding them with only certain kind of data, making them disregard some of the available data, or even choosing not to gather other types of data at all. This is the reason we have expert systems, or hybrid expert systems that are good at recognizing cats but are terrible at playing chess (and vice versa).

When dealing with moral scenarios AI entities will also face this limitation. In order to increase comprehensiveness of moral calculus they would need to take into account an increasing amount of observables. However, this would threaten to hit the upper boundary of an employed system. When it does, the system will have to simplify by disregarding some observables. This can open contentious moral issues of bias and discrimination: which observables can be disregarded, and which ones must be regarded? A possible solution is the *moral veil of ignorance*, whereby a system specifically disregards additional available information and chooses the course of (in)action (for a more detailed discussion see 2.3.1 Substantial (ethical) implications in Chapter V. Discussion).

Even for simple scenarios (e.g. trolley problems) the information AI entities use to reason will have to be a simplified version of the real world. For instance, the classical trolley problem (kill one person to save five) can look rather simple, and the solution trivial if we follow certain ethical theories. Consequentialism would typically answer positively, while deontology negatively.

However, what if that one person is a genius that was on his way to publish the cure for AIDS, but at the same time carries a dangerous pathogen that can potentially make a new Great Plague, while at the same time is the sole family breadwinner and takes care of 5 children, of which one is the potential next dictator…

As we can see, getting into too much detail creates a combinatorial explosion and renders reasoning impossible, potentially even for AI entities. Increased capacity for calculation might help, but will never eradicate practical reliance on heuristics. Hence, even AI entities will always have some kind of bias (except maybe for very simple scenarios).

3.4.3 Goals

In our everyday life we come across many different kinds of systems. Some of them obviously or even explicitly aim for something, while for others it is harder to make out if they aim for anything at all.

We have seen above that systems have a particular existence, a unified, integrated and unique whole that makes them a separate Being. They also have an internal structure, which has certain resistance to external pressure. All systems resist (unwelcome) change in their Being.

If we need to describe a single 'aim' that all systems strive for, it would be *resistance to change of Being*—or in other words, inertia of Being (which should not be taken as resistance to change in general—see the footnote[37]).

However, not all are successful in this endeavor. Earlier in 2.1.4 Panpsychism and in 3.3.3 Injury and destruction of Being I commented on the futility of total conservation of Being. The reason is the inherent stochasticity ('chaos'/entropy) of a system's environment that is ineradicable. Only systems that successfully *adapt* to this state of matters can (expect to) preserve their personal continuum through time and space.

In any case, all systems adapt to change. As Lotfi Zadeh is keen to say, "... all systems are adaptive, and the real question is what they are adaptive to and to what extent." (Zadeh, cited by (Klir, 2001; p. 171)). This adaptation to change can be an explicit or implicit internal striving of a system, and thus a *goal*.

All goals come by the nature of the system. Implicit goals are such that can be revealed as a striving of a system that is not explicitly encoded in its cause-effect repertoire. For example a stone, a mountain and a computer all resist change to their structure, yet this is (probably) not explicitly encoded in them. In contrast, biological systems and some synthetic systems can resist change or pursue other goals explicitly i.e. when a stock trading machine has the explicit goal of increasing the ROI margin by 5% from the current state. When goals are implicit they are named *instruments*, and when they are explicit they are named simply: *(explicit) goals*. Most goals are instrumental and subordinate to other goals, but some are an end in themselves. These last ones are titled *imperatives* (primary goals; see also 4.3.2 The moral imperatives).

Systems can also strive for other states of matters besides adaptation to change. For example, a human can strive for success in life, an algorithm can strive to fulfill its purpose and finish execution, an animal can strive to escape a predator or find enough food so that it does not feel hungry, a virus or a bacterium can strive to infect a host and multiply, a running engine of a car can strive to continue running, *etcetera*. Goals are one of the major focuses of cybernetics, as the following citation from Von Foerster testifies:

> "Here is Norbert Wiener, who re-introduced the term "Cybernetics" into scientific discourse. He observed, "The behavior of such systems may be interpreted as directed toward the attainment of a goal." That is, it looks as if these systems pursued a purpose!" (Von Foerster, 2003; p. 287).

What are goals

But what are goals, exactly? Here follows my definition of goals that I will use throughout this text:

> Goals $=_{\text{def.}}$ state of matters in the internal and external world that a system wants to see become real.

Thus goals are 'states of matters' that are yet *not* real. In a sense goals are imaginary models of the future.

[37] We should be careful not to interpret this as resistance to change in general. Change of Being means change of the structure that either breaks or significantly modified the continuum of Being throughout time and space. This can happen by e.g. change of structure that results in the change (or even deletion) of the goals of a system, change in structure so that identity is broken, and similar. However, if a system's Being specifically aims for change in itself, when this is achieved it is not a 'change of Being', it is exactly its opposite: the ***completion*** of Being (see also 4.2.1 The Good).

As in any model, goals are also comprised of observables. Let's recall that observables are *interpreted typed variables* that (aim to) represent some features of a system under consideration (see 2.1.6 Method of Levels of Abstraction). For example, a system in the present can have the observable of **color** with the value of *red*. That or another system may want to see the value of that observable change to *green* (as can happen with traffic light software). This is precisely a *goal* of that system.

Goals are typically described as limitations, and are always "in the eyes of a cognitive agent" (Klir, 2001; p. 171). Klir further describes goals as

> "defined in terms of some specific restriction of the systemhood properties that a cognitive agent dealing with the system considers desirable under given circumstances. Some examples of desirable goals are: keeping an output variable of a system within a specific and usually small range of values (point regulation); restricting the state transitions of a system to a specific cycle of states (path regulation); keeping a specific external behavior of a structure system invariant under some changes (malfunctions) in its elements (self-correction); acquiring in an autonomous way (through the regular operation of a system, with no specific interferences from outside) a particular spatial, temporal, or functional relationship (self-organization)". (Klir, 2001; p. 171).

We can see that goals are particular limitations in either types or values of observables that systems would like to see become real. For example, a system might aim for a traffic light to change to or remain 'green', or its body temperature to remain in the interval between 36.6 – 37 °C, or its partner to remain the one named 'Tina', or the number of its children to remain in the range between 2 and 6, or the state of 'war' in the territory it inhabits to remain at steady 0. These can all be described as intervals of values: ['green'], [36.6 – 37], ['Tina'], [2, 6], [0]. As I mentioned before, even discrete (e.g. Boolean) values can be described with intervals that do not include ratios i.e. that only take boundary or discrete values into consideration: [true, false] or [blue, green, red].

Some typical goals that systems might have I already mentioned. For example, resistance to change of Being is one. It also translates to *conservation of personal continuum* (CPC) when describing it through time, for the **future** state of matters. This is why I include CPC as one of the two primary moral imperatives of systems (see 4.3.2 The moral imperatives).

Formal representation

To formally represent goals, we can take all the aforementioned and represent it symbolically with simple mathematical and set theory.

Goals are states of matters that a system wants to see become real (in the future). These states of matters ought become real in the world. Hence, we are discussing about the world of the future.

We can take the letter *W* to describe the world, W_f to describe the world of the future, and $W_f t_n$ to symbolize the world of the future at time t_n. Worlds are sets of states of matters *S*. We can further symbolize states of matters as sets (ordered pairs) of observables *O* (who are interpreted (*Int*) typed (*T*) variables (*x*)), where each observable has a particular set of limitations *L* at $W_f t_n$. Limitations are described as intervals between two values (that belong within the observable's type), Cyrillic *а* and *б* (that transliterate to Latin *a* and *b*). Hence,

$$W_f t_n = \{ S_1 \dots S_n \}$$

$$S_n = \{ (O_1 \dots O_n), (L_1 \dots L_n) \}$$

$$O_n = \{ x, T, Int \}$$

$$L_n = \{ [а, б] \mid а, б \in T \}$$

This is a very simple (meta)representation that can be upgraded with some key elements from decision theory (e.g. strict $\prec$ and weak preference $\preccurlyeq$), but there is no need to dive that deep for now.

3.5 Systemic resources

In order to dynamically function all systems need resources. Here I will include a modified definition of resources from a published article of mine (Dameski, 2018):

> Resource = $_{\text{def.}}$ A part of the world which a system can use instrumentally to pursue its goals. This includes both traditional ones such as raw materials, energy source(s), food, water, minerals and similar; but also time, situations, rules, other systems and their parts, and anything else of utility.

So, a resource is anything in the world that helps systems in their pursuit of goals. Since all entities are systems, most traditional resources are also systems. This means that they can be described with the method of LoAs through observables and their relations.

For example, 20 liters of water, 2 cars, and €300 and $500 can be described in the following way:

Resource	Observable	Value
(available) water	volume	20 l
(available) cars	identifier	Fiat Multipla 1st generation Yugo Coral 55 hp 1989
(available) money	currency	€300 $500

Non-traditional resources are features or non-systemic phenomena, but if they are perceivable and/or measurable they can also be described through LoAs:

Resource (LoA)	Observable	Value
(available/ estimated/etc.) time in the future	seconds	14,250
(available) weather in the future	State at time t_n	sunny
(available) law	Provisions (of the Civil Code)	Article 5: *[text of provision]*
(available) military power under command	Soldiers (individuals)	541

Different LoAs can be conjoined or even nested inside other LoAs to increase the power of the model. For example, in the last table in the LoA **weather in the future** we can nest **time in the future** so that we don't specify an observable **State at time t_n** but simply **State**. With this we will be creating what Floridi describes as *interfaces* (Floridi, 2013; p. 32).

3.6 Processes

Processes are phenomena, transformations ('happenings') in the world that cause changes in the values of observables and/or in their relations. All actions, reactions, and even inactions (where actions were possible and relevant) are processes. By extension, moral processes are a subset of systemic processes.

3.6.1 Systemic action

As mentioned above, actions are processes that cause changes in the values of observables and/or their relations.

Systems use actions to pursue their goals. Recall that goals are particular limitations in either types or values of observables that systems would like to see become real (see 3.4.3 Goals above). Systems attempt to choose actions that make certain wanted or expected changes happen. This can be attainment of resources, helping or harming other systems (e.g. by improving or decreasing their *quality of life*), sending a message (e.g. by changing the value of some system's data observable), and similar.

Actions of systems can also be described through LoAs. This gives the possibility to describe the (envisaged or performed) changes over resources. For instance, an action of the type **movement** from one to another point in space and time could result in a negative change of available resource **gasoline** of 2 liters.

Action (LoA)	Observable	Type	Value	
movement	time	t_n	t_0	t_1
	position in space	coordinates (x, y)	12, 30	15, 15
	amount of resource (gasoline)	liters	15 l	13 l
	change of resource (gasoline)	Δliters	Δ = 0	Δ = -2
	satisfaction	feeling	good	better

The more observables we add the more precision we have available with which we can take an action into account. However, the more observables the more cognitive or computing power is required.

3.6.2 Systemic reaction

Systems exist in a dynamic world. While attempting to pursue their goals they are subject to influences and pressures from other systems or features of the universe (e.g. entropy). In order to be more successful they need to react to these phenomena. Reaction is a very important subject of cybernetics (being a study of communication and control in systems; see 3.4.1 Reasoning and cybernetics). I already mentioned in 3.3.3 Injury and destruction of Being that Beings need to maintain a dynamic order—in contrast to a static, rigid one.

Systemic reaction is also very important for morality. Through an accumulation of actions and reactions systems establish balance points that have predictable effects on QoL, and these balance points are moral rules (see also (Dameski, 2018)), especially when explicitly recognized.

In regards of representation systemic reactions are, in essence, actions. To represent them we can use the same conceptual tools that we use with actions. The LoAs and interfaces we use can include some additional observables that describe to which action, system or phenomena is the reaction a response.

4 Towards Ethics of Systems

The time has finally come to formulate a coherent (meta)ethical framework that can help manage moral scenarios where systems in general—and by extension AI systems—are participants.

But first, a word or two about the theory itself. Ethics of Systems is what Floridi calls *macroethics* (Floridi, 2013; p. 25), or even better described as *holoethics* by Fultot (2016).

A macroethics according to Floridi is a "theoretical, field-independent, applicable ethics" (Floridi, 2013; p. 25). That is to say, it is based on a particular theory and aims to describe, prescribe and govern all morally-relevant phenomena everywhere and at all times. Fultot notes that since such ethics (i.e. Floridi's Ethics of Information) are concerned with "… behavior **holistically** and globally as opposed to locally and individually" [boldface mine] (Fultot, 2016), it should borrow the 'holo-' from 'holistically' and be called *holoethics*.

Ethics of Systems theory aims the exact same thing. It aims to describe, prescribe and govern all morally-relevant phenomena and states of matters everywhere and at all times, building upon the (meta)ethical foundations of systems theory, philosophy and ethics of information, and classic and upcoming ethical theories.

It is without doubt that this is an endeavor that might span years and even decades—a *magnus opus* of a sort. Therefore I reserve to present more detailed explorations (and revisions) of the theory in the future, and stick to the foundations and necessary details in regards to AI ethics here, in this thesis.

Finally, some commentary about the label choice. The framework and theory is titled Ethics of **Systems** because it is based on systems theory for its ontological and epistemological foundations. The main subject of study are systems, their existence, states, behavior, and everything else which is morally-relevant. Even though it is heavily inspired and influenced by Floridi's Ethics of Information[38], it is not a part of IE. The main reason is that it considers information as an emergent systemic property. This is the reason why the systemic LoA (level of abstraction) comes conceptually before the informational LoA.

4.1 The Ethics of Systems Interface

If we want to perform an exhaustive and methodical study over ethical perspectives regarding systems we need to have the right methodological tools. I am using the method of Levels of Abstraction (LoAs) already described in section 2.1.6 Method of Levels of Abstraction above.

At times there will be a need of conjoined or nested LoAs depending on the (class of) moral scenario under study. This is why I am designing an *interface*, as a comprehensive conceptualization that contains all the LoAs, their observables, their types, interpretations, and relations that are needed by this Ethics of Systems Framework. Of course, each and every component LoA is independent, and does not have to be embedded in the Interface to serve some function.

The Ethics of Systems Interface is comprised by the following major LoAs: **moral scenario, moral entity**, **moral process**, **moral theory**. Some further (sub)LoAs can be, for example, **resource**, **environment**, **entropy**, and **miscellaneous**; but are not included in the table below. The default values of observables are specified as *unknown / undefined*. This list is by no means final. As the Ethics of

[38] Here I must mention that the idea of Ethics of Systems has come to me long before I got acquainted with Floridi's work. It originally sprang out from an intense and prolonged philosophico-ethical discussion with a bunch of very dear friends of mine, amongst which one of them, Filip, boldly proclaimed: "The Good is life!" This got me pondering for months and even years, before eventually resulting in Ethics of Systems. I can't thank you enough, my Filip and my friends.

Systems Framework is further developed there will no doubt be additions, modifications or deletions of parts of the Interface.

In a tabular representation:

Table 5: The Ethics of Systems Interface (integral version)

LoA	Observable	Type	Interpretation
[Moral scenario]	Class	Class identifier: *name of class of moral scenario*	A class identifier (*name of class of moral scenario*). Can contain multiple values.
	ID	Personal identifier: *name of moral scenario*	A personal identifier (*name of moral scenario*) of a particular moral scenario. Can contain multiple values (but good practice is to settle for one or few).
	Moral entity	Nested LoA [Moral entity]	In a particular scenario the moral entities are nested from the LoA [Moral entity]. The nested LoA is uniquely instantiated for each (class of) moral entity.
	Moral process	Nested LoA [Moral process]	In a particular scenario the moral processes are nested from the LoA [Moral process]. The nested LoA is instantiated for each moral process.
	Moral theory	Nested LoA [Moral theory] $T = (M_{1 \dots n},\ R_c)$	Nested and instantiated set of moral rules formed according to the specific axiology of the theory (see [Moral theory] below).
	Time	Placeholders: t_n or time units: *yyyy-mm-dd-hh.mm.ss*	Time can have some default placeholders, such as t_c, t_f (the current or a future time frame).
	Time resolution	Ratio *step* : *time* 1 : t_{n+x} 1 : yyyy-mm-dd-hh.mm.ss + x	The resolution (granularity) of time steps under study. It is a ratio between one step and one jump between discrete time steps. Under ‘time steps’ are understood the smallest discrete units of time between which the moral scenario model can represent change.
[Moral entity]	Class	Class identifier: *name of class of moral entity*	A class identifier (*name of class of moral entity*). Can contain multiple values.
	ID	Personal identifier: *name*; *name in scenario*	A personal identifier (*name*) of a particular moral entity. When nested in a Moral scenario, the ID includes also *name in*

			scenario. Can contain multiple values (but good practice is to settle for one or few).
	QoL	*CPC* x *APG* (arithmetical product) $QoL \in \mathbb{R}$ & $QoL \in [0, 1]$	The state of Quality of Life is the product of the states of Conservation of Personal Continuum and Achievement of Personal Goals. Since they interact and influence each other, their product is arithmetic. QoL can only hold value (rational number) within an interval of 0 and 1.
	CPC (Conservation of Personal Continuum)	$CPC \in \mathbb{R}$ & $CPC \in [0, 1]$	CPC can only hold value (rational number) within an interval of 0 and 1.
	APG (Achievement of Personal Goals)	$APG \in \mathbb{R}$ & $APG \in [0, 1]$	APG can only hold value (rational number) within an interval of 0 and 1.
	Moral respect (reputation)	$O_m \in [0, 1]$ $O_m > 0$ $O_m = 0.5$	Moral respect (O_m) is measure that reflects the reputation of an entity in a scenario. It ought be taken in consideration alongside Moral status (S_m; see below) to determine overridability. Moral respect belongs to an interval between 1 and 0, and cannot reach zero (but can approach it). If not specified otherwise by class, the default value is 0.5 (neither respectable nor not respectable).
	Moral status	$S_m \in [0, 1]$ $S_m > 0$	Moral status (S_m) can never reach zero. It belongs to an interval between 1 and 0, and has initial state determined by the class of [Moral entity].
[Moral process]	Class	Class identifier: *name of class of moral process*	A class identifier (*name of class of moral process*). Can contain multiple values.
	ID	Personal identifier: *name in scenario*	A personal identifier (*name in scenario*) of a particular moral process in the particular scenario. Can contain multiple values (but good practice is to settle for one or few).

	Time of availability	t_a	Time of availability specifies the time frame of availability for consideration of the process by the moral scenario, moral entities, and moral theory.
	Time of execution	t_e	Time of execution is a particular time frame of the moral scenario at which a particular process' effects start to take place. In simple scenarios the time of execution is the same with the time of availability observable.
	Agent	Nested observable: $_{\text{[Moral entity]}\rightarrow}$Class, ID	
	Patient	Nested observable: $_{\text{[Moral entity]}\rightarrow}$Class, ID	
	Effect	Change in values of observables: *observable* ($_{\textit{[Moral entity]}\rightarrow}$*Class*, *ID*, *observable name*), *change* (*Δ*) $\Delta = \{\ (O_x, \Delta_O) \ldots (O_n, \Delta_n)\ \}$ $\Delta \in \mathbb{R}$	The total effect of the moral process is represented as a change (Δ) of each affected observable's value, that belongs to a particular moral entity, by a certain amount. This amount is any real number, or can be specified formulaically.
	Effect duration	Time steps ($_{\text{[Moral scenario]}\rightarrow}$Time resolution)	This observable tracks how much time it takes from the initiation of the effect to its conclusion[39].
	Effect on QoL	$\Delta QoL \in \mathbb{R}$	This is the effect of the moral process on QoL for a particular time frame (if needed for granularity).
	Cumulative effect on QoL	$\Delta QoL_c = \sum \Delta QoLt_{0\ldots n}$ $\Delta QoL_c \in \mathbb{R}$	This is the cumulative effect of the moral process on QoL for all time.
	Rule pertinence	$_{\textit{[Moral theory]}\rightarrow}$*moral rule*, *pertinence* $M_{1\ldots n}$, $Mp_{1\ldots n}$ $Mp \in [-1, 1]$ $Mp \in \mathbb{R}$	This observable tracks to which moral rule each moral process pertains, and how i.e. whether it satisfies the rule in the positive, neutral, or negative direction, and how much. The satisfaction of the rule thus lies on an interval between -1 and 1. When an action fully supports a rule, Mp receives a positive value bounded by 1. For an action that acts completely against a rule, Mp receives a negative value bounded

[39] Further improvements can be in granulation i.e. how is the effect spread over each time step, whether evenly or not.

			by -1. And finally, for an action that neither satisfies nor dissatisfies a rule (i.e. one that does not pertain to a rule), Mp receives a value of 0.
	Choice value	$Vc \in \mathbb{R}$	Choice value (Vc) is the value that each available process gets assigned by executing a particular moral theory. Choice value can be assigned by the scenario by and for itself (and its components); and/or by any participating entity for themselves and/or for other entities or for the scenario.
[Moral theory]	Class	Class identifier: *name of class of moral theory*	A class identifier (*name of class of moral theory*). Can contain multiple values.
	ID	Personal identifier: *name of moral theory*	A personal identifier (*name of moral theory*) of a particular moral scenario. Can contain multiple values (but good practice is to settle for one or few).
	Moral theory	A set T, comprised of sets M and R_c: $T = (M_{1\dots n},\ R_c)$	A moral theory is a set comprised of two sets: the set of all moral rules within that theory, and the set of their relations (according to criterion *c*).
	Moral rule	Set of ordered pairs of *goals*, and their *importance* (interval): $M_{1\dots n} = (G_{1\dots n}, I_{1\dots n})$ $I \in [0, 1]$ The set of all rules: $M = \{ M_1, \dots , M_n \}$	A specification of what is permitted, forbidden, desired, and/or required. These are all defined as *goals* (sets of states of matters in the future; see 3.4.3 Goals) of the moral theory that systems ought to accept as their own, and order them according to *importance* (see Relation below).
	Relation	A subset of all possible relations *R* in *T*, according to criterion *c*[40]: $R_c \subseteq T \times T$ Criterion *c* can describe 5 different relations: a) equivalence: $M_x \sim M_y$ b) compatibility: $M_x \approx M_y$ c) partial ordering[41]: $M_x \preccurlyeq M_y$ d) strict ordering[41]: $M_x \prec M_y$	A relation is a mathematically-described connection between two moral rules (M_x, M_y) from the moral theory set (T). This relation is selected as a subset of all possible relations according to a criterion (*c*).

[40] For more detail see Klir (2001; p. 13).

		e) other relation Ordering in c) and d) is being done according to *importance* (I).	

Table legend:

As we can see above, **[Bold text in square brackets]** is used to show and refer to the LoAs. Observables are titled with a Capital first letter (i.e. Moral rule). At times, when we need to refer to a particular observable and/or its symbol, we can use smaller font to explicate the whole chain e.g. [moral process]→Choice value→Vc. The symbol ":=" means "is assigned the value or property of".

Please note that this is the *basic* EoS Interface, because it includes the fundamental observables required for any moral scenario. However, it can always be upgraded by adding new observables where appropriate. One example would be adding the observable **Goals** for the moral entity, whereby all the goals of the entity would be listed separately (instead of including them under APG and CPC).

Another important thing to note is that any of the observables and LoAs can be nested inside each other[42], but only if appropriate. This does not imply that they will have the same *values* of the observables, however. An example would be nesting the LoA **[Moral theory]** within both the **[Moral scenario]** and all separate **[Moral entity]** instances of that scenario. Therefore, the separate **[Moral entity]** instances will have their own moral theories to consult, or, if allowed, can also consult the moral theory of the scenario. Additionally, rules in a **[Moral theory**] instance can govern which **[Moral theory]** takes precedence during consultation and application, in order to resolve conflicts between moral theories.

As noted before, this only is to be done if appropriate. For example, it mostly makes sense to nest **[Moral process]** within **[Moral scenario]**, in contrast to nesting it within any instance of **[Moral entity]** (with exceptions). Another example would be where having subjective moral theories for moral entities is not important, so it makes no sense to nest **[Moral theory]** within **[Moral entity]**. Design choices need to be taken wisely in order to decrease complexity of the design while increase the capacity of computational and formal representability.

4.2 Axiology

There is no thorough ethical theory or framework without proper axiology (value theory). As I have discussed already in the literature review, all ethical theories consider certain state of matters as valuable, others as non-valuable (irrelevant), and yet others as negatively valuable (see Chapter II. Literature review 3.1 Value theory; and also (Hurka, 2006)).

[41] Weak and strong ordering is decided according to *preference* (e.g. moral rule M_x is preferred to M_y because of criterion c). Preference is denoted with $\preccurlyeq$ and $\prec$, instead of with classic mathematical ordering symbols $\leq$ and $<$. See Steele and Stefánsson (2016) for more detail.

[42] When a level of abstraction is nested within another (e.g. **[Moral entity]** nested in a **[Moral scenario]**), it can be treated as both a LoA and an observable (that contains other observables within itself). The preferred approach depends on the implementation.

The Ethics of Systems Framework is, on first sight[43], best suited to represent consequentialist ethical theories. However, it is fully capable of representing deontological theories by using *relations* between rules. Therefore it can represent ethical theories that both support commensurability and incommensurability.

Ethics of Systems has its own conceptualizations of the (morally) Good, Bad, Evil, and neutral. As is typical with sophisticated theoretical frameworks, there is an interplay between all the aforementioned. All of them are multifaceted, and dependent on the rest regarding context.

Earlier in Chapter II. Literature review in 3.1.1 The value spectrum I have included a graphical representation of axiology. Here I will include the Ethics of Systems value spectrum, which can help in having a more intuitive image of the Framework's value system.

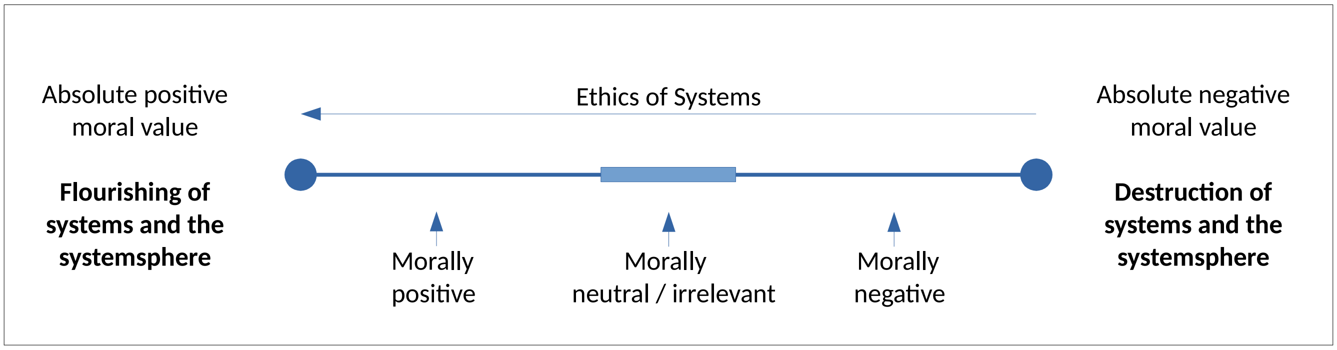
Absolute positive moral value
Flourishing of systems and the systemsphere
Ethics of Systems
Absolute negative moral value
Destruction of systems and the systemsphere
Morally positive
Morally neutral / irrelevant
Morally negative

Illustration 5: The Ethics of Systems value spectrum

As we can see, morally positive processes and states of matters are located and/or move to the left while morally negative ones move to the right. There is a certain space in the middle for the morally neutral and irrelevant. Beings and existence are always in the morally positive, while processes can be marked everywhere on the spectrum.

In contrast to Floridi's IE, Ethics of Systems admits that both actions *and* states of matters can take a place on the value spectrum i.e. be good, bad, evil, tragic, neutral, or irrelevant. We will see immediately below that (besides actions) some states of matters such as flourishing, existence and goals are morally positive; destructive states of matters are morally negative; some even out as morally neutral; and some have no moral relevance.

4.2.1 The Good

From the value spectrum I included above we can see that absolute positive moral value is 'reserved' for flourishing of systems *and* the systemsphere. That is to say, this is the highest possible Good. Note that the statement includes <u>and</u> (the systemsphere), not <u>or</u>. There is a very good reason for this that will be explored in a short while below.

The Good, as defined in Appendix II. Key concepts, is what is considered as valuable from the perspective of ethics and morality. In Ethics of Systems the Good is the flourishing of systems and the systemsphere.

The Good $=_{\text{def.}}$ A state of matters where systemic Being is able to, and does, flourish.

[43] In general the broad distinction is between states of matters that are commensurable and incommensurable. That is to say, for some theories differing states of matters are comparable, for some are not. Thus we have the distinction between what is good and what is right. Both approaches have advantages and disadvantages. In Ethics of Systems both these perspectives are taken into account, and they can both be represented within the framework.

But what is this *flourishing* we are talking about? In virtue ethics we meet the concept of *eudaimonia* (comprised of *eu-* [good] and *daimon* [spirit]). According to Aristotle,

> "Every science, investigation or action aims at some good. Such goods exist in a hierarchy: the lesser goods are instrumental in seeking the higher goods, but many things are good in and of themselves. ... The highest good will be the final goal of purposeful striving, something good for its own sake (...). This final good for human beings is *eudaimonia* (happiness), which is always an end in itself ... The goodness (*arete*) of anything—including human beings—resides in its proper function (*ergon*)" (Johnston, 2014).

We can notice several key elements here. Everything which is morally relevant is a striving, an aim at some good, and thus is a *goal*. The final, or ultimate, goal towards which all other goals work is eudaimonia (flourishing[44]). Additionally, whether something is good or bad depends on its *proper* function. If something does not function well (i.e. towards the Good), it is not good.

From the last statement we can extract the following argument: since the Good is the ultimate goal, and the ultimate goal is flourishing, flourishing is the Good. Additionally, all actions and states of matters that function towards the Good are good; and those that fail to function in this way are not good. Finally, flourishing, being a goal itself, is a state of matters itself—the ultimate morally positive state of matters[45].

We are also discussing about the *ultimate* Good. It is an envisaged final state of matters where the moral process is *completed*. That would imply that until this state is achieved, the moral process is *not completed* (Annas, 2006; p. 521). The purpose of moral (in)actions and deliberations is to bring a Being's moral process into completion. We can describe this final state of matters as *perfection*. Therefore, the Good is a perfect (moral) state of matters, and a perfect Being.

That the Good can be defined as a perfectionist account of (the systemic) Being is nothing strange. Brink (2006; p. 389) argues that "what is good for someone is what his idealized self would want his (nonidealized) self to want". He then goes on further to say that "There is a venerable perfectionist tradition, common to Aristotle, Mill, and T. H. Green, among others, that identifies a person's good with the perfection of her nature" (Brink, 2006; p. 391). Hurka also comments on perfectionist approaches, saying

> "Some more ambitious approaches try to unify all the perfectionist goods. One appeals to the concept of human nature, which in different formulations it takes to consist in those properties essential to humans, distinctive of them, or essential and distinctive (Hurka, 1993, ch. 2). Its central idea is that the good in a human's life consists in the full development of whatever is fundamental to human nature; it is often generalized to hold that the good of any natural thing consists in developing its nature. This view can generate different particular values, depending on which properties it takes to constitute human nature" Hurka (2006; p. 365).

4.2.1.1 UNIFORMITY OF BEING AS THE GOOD, AND INTRINSIC VALUE

Good is also about existence **now**, not just in the future. In a sense existence now is even more important than existence in the future, and maybe the *only* important state of matters. This is because systems have goals for the reason that they want certain states of matters to be true now. If these cannot be true now, the second-best thing is to envisage such states of matters (and ways to reach them) for *another now* in the future. If and when they do become true, they will become true in the what then will be *now*.

[44] The author uses *happiness* as translation for eudaimonia; but recently the word *flourishing* is taken as more suitable (Annas, 2006; p. 520)

[45] We should not, however, confuse the seemingly static term 'state of matters' with a static situation it is describing. Flourishing need not be, and indeed typically is not, a static situation. Flourishing for some systems may also represent a constant dynamic situation in which it consistently moves towards achieving multiple goals, but never reaches a state where there are unfinished endeavors left.

And what is now for any system, is existence. The systemic Being is existence. To exist is to be alive. On the other hand, if a system ceased to exist there is no more 'now' for it. This is why—and also in what way—now and existence are substantially connected. The fundamental definition of life according to James Grier Miller is maintaining “a steady state of negative entropy over a significant period” (Miller, 2001; p. 714). Every system is something, and that something is different from everything else, including non-existence. This means that every system is unique, uniform and integrated. Hence an entity has a so-called *uniformity of Being* (Floridi, 2013; p. 65).

However, in stark contrast to Floridi’s IE, as already discussed before, Ethics of Systems allows for contradiction within systems[46]. One example is a system that is locally integrated which makes it a closed system; but at the same time integrated into systems of higher orders and eventually into the systemsphere, which makes it an open system. The resolution of this contradiction is the understanding that any and all systems (except maybe the systemsphere itself) are always semi-open/closed. Whereas, in Floridi’s IE,

> “[A]n entity is a consistent packet of information, that is, an item that contains no contradiction in itself and can be named or denoted in an information process. A contradiction, when directly and positively used (i.e. not used at a meta-theoretical level, or just mentioned), is an instance of total negation of information, i.e. a mark left where all information has been completely erased, a scratch in the fabric of the infosphere. Since an information process positively involving a contradiction ends up being itself a source of contradiction, it is also a case of total negation, an information black hole, as it were” (Floridi, 2013; p. 65).

To exist is to be alive. Every and any system ‘wants’ to be alive, whereby being alive means achieving personal goals and conserving personal continuum at least until those goals are achieved. Being alive is the essence of any and all systems.

Intrinsic value and moral respect

Therefore, life (existence) itself is morally valuable, primarily personally for the system itself (moral egoism; see 1.1.3.1 Fitting Attitudes, agent-relative and agent-neutral ethics, and Universalizable egoism below). A system has the ‘right’ to have its existence and its goals by the virtue of its uniqueness and existence, which is why these also have moral value (i.e. are morally relevant). This ‘right’ to exist is a Being’s *intrinsic* value.

For a system, its existence and its goals have *prima facie* systemic value that arises out of the perfectionist account I mentioned above (similarly, C. S. Lewis defines moral bad as ‘broken good’; see (Lewis, 2017; p. 69; Macedonian translation)). In order to have goals at all, and then go about to achieve them, a system has to exist in the first place (even more so since part of its existence are its goals). Therefore, existence *now*—as the system is or as it wants to be—has **intrinsic value**, as do the goals themselves. This intrinsic value can also be named **moral dignity**. Moral dignity deserves **moral respect** from other moral entities. It is important to note, however, that moral respect for each Being is overridable according to context (see below).

[46] The discussion on contradiction is pretty important, so it bears some more commentary here. Ethics of Systems allows for each system to contain contradiction in itself. This does not mean that the contradictory aspects (‘natures’) are ‘activated’ *at the same time*, or if they are, that they are *equally powerful*. For example, the electrochemical systems of the heart or of neurons work in contradictions—the heart’s muscle and nervous command works in a circular fashion, whereby they cycle between activated/inactivated *throughout time*. However, the same components are never both activated and inactivated at the same time (which would mean a heart mechanism failure—if possible at all) i.e. the muscle pumps in and then pumps out only to pump in again etc. This is an example of temporal distance between contradictions (non-simultaneity). On the other hand, an example of simultaneous but differential power would be compressed gas into a canister. Here we have gas that pushes to break out, but if the canister is *stronger*, the whole system will stay intact and continue to function as expected. However, if the pressure overcomes the canister’s strength the system is destroyed (it *totally* contradicted itself with *equal power of contradictions* and *at that moment*).

Additionally, a system wants to be 'perfect' (in some future *'now'*), and assumes it will perfect itself by achieving its goals. Therefore, existence *in the future* until all goals are achieved also has intrinsic value. This is why the moral imperatives of Conservation of Personal Continuum and Achievement of Personal Goals are so tightly intertwined, and a system's Quality of Life (QoL) is solely their product (see 4.3.2 The moral imperatives below).

It is important to note that Moore expressed skepticism towards the claim that what exist is good, seeing it as a naturalistic fallacy perpetuated by naturalistic philosophers and metaphysicians. He thus explicitly states that it still remains a distinct and different question whether what exists is good (Baldwin, 2010; p. 288). Ethics of Systems, as well as IE, answer this question in the positive.

A final comment is in order on the linguistic choice I espouse below. As existence is synonymous with Being and with life, this synonymy is the reason why I name the measure for (degree of perfect and complete) Being of a system as Quality of **Life** (see below).

Now we need to explore what flourishing means for systems and for the systemsphere.

The Good of systems

First, let's work on systems. As we have seen in 3.4.3 Goals above, systems have certain states of matters, entitled *goals*, that they want to see come true in the world. If their goals take place we can speak of *completion*—that is to say, the system has been completed (has become perfect according to internal and/or external account).

To reflect the state of completion, I am introducing the measure entitled **Quality of Life** (QoL). It is a product of the two most important goals of a system, the **Conservation of Personal Continuum** (CPC) and the **Achievement of Personal Goals** (APG); who are described as moral imperatives (see 4.3.2 The moral imperatives below).

If we see it as binary state, the maximal QoL as a final state of matters can either be completed or not. Hence we can describe it symbolically with either *achieved/completed/true/1*; and *not achieved/incomplete/false/0*. But goals in general can be traced in their completeness, between 1 (complete) and 0 (incomplete). Additionally, since maximal QoL is a goal, it can also be traced differentially, as a difference between the current state of matters and the envisaged final and perfect one. Hence we can describe the current state of matters as laying somewhere in the closed interval between 1 and 0. QoL can hold the value of, in example, 0.8, 0.2, and any other rational number between 1 and 0.

The general aim is to increase QoL until it achieves value of 1. This, or close-by, or steadily moving towards it, is when a system would be flourishing. Or, inversely, the general aim is to avoid decreasing QoL by avoiding actions and states of matters that have such deprecating effect on it. Additionally, I already mentioned in 3.3.3 Injury and destruction of Being that in a world where entropy is a constant reality, systems at least ought to (primarily) aim at conserving their existence; which would translate in aiming to conserve the current value of QoL.

With this we can also define what Good (moral) process and state of matters is:

> Good process / state of matters = $_{\text{def.}}$ what contributes to an increased, or at least conserved, Quality of Life (QoL) of systems.

QoL is not a standalone goal, but is a reflection and thus the product of the two moral imperatives, CPC and APG (which are also represented on an interval between 1 and 0). In order to obtain the value of QoL we need to multiply CPC and APG (see 4.3.2.1 Quality of Life). That is, QoL = CPC x APG.

The Good of the systemsphere

Now let’s turn to the whole systemsphere. In 3.1.4 The systemsphere I defined the systemsphere as a set that contains every system (entity) that exists, have existed, and will exist. Besides being a set of all systems, the systemsphere is also an integration of all systems, thus being a system (entity) itself. A system of all systems—a totality of Being.

I commented on something similar in 2.1.4.1 Soft exclusion in 2.1.4 Panpsychism above that integration (of consciousness, but also in general) can be local, superlocal and global—*in parallel*. All separate systems are locally integrated, yet at the same time they are integrated in systems of higher order, and finally within the systemsphere.

As the systemsphere is also a system, its own (universal) flourishing also has absolute positive moral value for itself, and by extension, should also for all its components. Actions and states of matters inside the systemsphere ought to contribute to the flourishing of the whole systemsphere and all its systems, not just to some particular ones. Just like with any wise and mature collective, there is a balancing act between the Good for the collective and the Good for every individual member. Since the existence of *all* systems has intrinsic value, the value of the whole systemsphere is initially *incommensurable* to the value of each particular system; and the value of each system is initially incommensurable to the value of any other. For practical purposes they might be regarded as equal in value, but if we are following the theory there is another way in which we can determine how each systemic Being’s value compares to the value of any other (see below).

Therefore the intentional or negligent injury and destruction of a particular system is allowable (and hence its moral dignity overridable) only in defense of another system or the whole systemsphere. This defense has to be proportionate to the context, and serve as a deflector (disabling an unduly aggressive action of a system), deterrent (‘I will act in (self)defense if you attempt to injure my Being’) and as retribution, so that it is effective all while injuring the original offender’s Being in the least possible amount. When injury is introduced in proportionate (self)defense it is not moral bad.

4.2.1.2 Ethics of Systems’ four ethical principles

Floridi offered his basic ethical principles of Ethics of Information (see 2.2.2 The four ethical principles of IE above). Here I will offer a modified version of these that comply with Ethics of Systems.

Table 6: The four basic ethical principles of Ethics of Systems

0 Destructive entropy ought not to be caused in the systemsphere (null law)
1 Destructive entropy ought to be prevented in the systemsphere
2 Destructive entropy ought to be removed from the systemsphere
3 The flourishing of systems as well as of the whole systemsphere ought to be promoted by preserving, cultivating, and enriching their well-being

Similarly as in Floridi (2013; p. 71), these principles are listed in an increasing moral value, which can be traced by their number. They also ought to be respected in a particular manner. For a moral process to be approvable and its source (agent) praiseworthy, it ought to satisfy the combination of the null law and at least one other principle. In contrast, a moral process is increasingly less approvable and its source more blameworthy the lower is the number index of the specific principle that they fail to satisfy.

4.2.1.3 COMMENSURABILITY, AND OVERRIDABILITY OF MORAL RESPECT

Ethics of Systems' basic components (e.g. the EoS Interface, the four ethical principles, the moral entities, moral process, moral theory, moral scenario, etc.) are able to accommodate different ethical theories, which is one of the primary goals of this work. Different ethical theories have different axiologies, and this translates into considering different things as valuable, and/or the same things but in a different degree.

At times a comparison between the value of different elements of a moral scenario needs to be done. For example, when we are exploring trolley problems we need to calculate the value of the one person we might push in comparison to the value of the 5 persons tied on the railway. Our own participation also enters the calculus. It is a different calculation when someone else is pushing the person before the trolley, and when we are doing the deed (see 3. Moral scenarios within Ethics of Systems in Chapter IV. Applying Ethics of Systems Framework to AI moral scenarios (the Ethics)).

Deontologists would typically state that the value of the different persons (or of different actions that ought to make a judgment on persons' value) are incommensurable (incomparable). Therefore we are not allowed to throw a person on the railway to save 5 other persons because a person's dignity is incommensurable to those 5 persons' dignity (unless that person decides to *sacrifice himself*, which is the basis of heroic ethics). Consequentialists might typically calculate that saving 5 persons is worth losing one (if we operate under the *veil of ignorance* and we don't want to have no idea what the persons did and what is their moral status at the time of decision-making). All these moral processes can be accommodated by Ethics of Systems by using relations between moral rules and rule pertinence values to determine their ordering (see 2 Moral theories within Ethics of Systems in Chapter IV. Applying Ethics of Systems Framework to AI moral scenarios (the Ethics)).

4.2.1.4 MORAL STATUS

But if we need to explore Ethics of Systems to the end by using its own principles, there is a way how to determine commensurability of Being. As was said before, Being has inherent moral value (moral dignity), and this moral value deserves **moral respect**. However, moral respect is *overridable*. That means that in certain contexts it can be partially or fully disregarded because something else's moral value is bigger. The question remains: how to determine the value of each Being? How do we know that a particular system's value is bigger or smaller than the other system's in that context?

The answer is by utilizing the concepts of **moral status** and **moral respect**. **Moral status** is the inherent value of a system—inherited from its class—which is then used in a computational function with **moral respect** to provide a commensurable moral value of a Being at a particular moment.

At conception moral status has value somewhere on an interval between 1 and 0. On the other hand, the value of moral respect can be specified by the class, but if it is not, its default value is 0.5 (neither respectable nor not respectable). The value of moral respect is affected by each good or bad action an entity performs, or good or bad state of matters it contributes towards becoming true, in the universe. Then, by computing a function between moral status and moral respect we arrive at the commensurable value for that Being at that particular moment. With it, we can compare a Being's commensurable value to that of another Being, and see which one is higher—and thus overrides the other Being's one at that particular moment.

Moral respect can be approaching but it can never reach zero (or negative) because all Beings, regardless of what they did, have some inherent value left that cannot be diminished. It can also never exceed 1, which is the maximal moral respect any moral entity can have. Similarly, moral status cannot reach 0, and neither can it exceed 1.

All systems have inherent value. However, initially (at conception) some systems which have greater capacity to do Good have greater moral status—determined by their class. This is because they have greater potential to follow the four ethical principles of Ethics of Systems (see above), and thus ought be helped in this by considering them more valuable. They are thus more worthy of efforts to help their Being be conserved and flourish, because it is assumed that they can and ought contribute more towards the Good in comparison.

After this initial state at conception, though, systems that **actually** contribute more towards the Good are assigned greater commensurable moral value in comparison to others. For instance, human beings have the same moral value at conception, but some of them become scientists, heroes and Martin Luther Kings, while some become Hitlers, Stalins, Maos, and guards in secret prisons, concentration camps and gulags. Obviously the moral value of the aforementioned is not equal or close to equal.

4.2.1.5 In conclusion of the Good

To round off the argument, the ultimate positive value (Good) is the flourishing of systems and the systemsphere. Flourishing is disturbed by the introduction of destructive entropy that threatens to destroy the capacity of a system to pursue its goals, and even destroy the system itself. This is moral bad, which I will explore immediately below.

4.2.2 The Bad and the Evil

As with any serious ethical theory, Ethics of Systems contains a definition of moral Bad and moral Evil. But before I continue I will use this opportunity to note that Evil is a subset of Bad, not a separate concept.

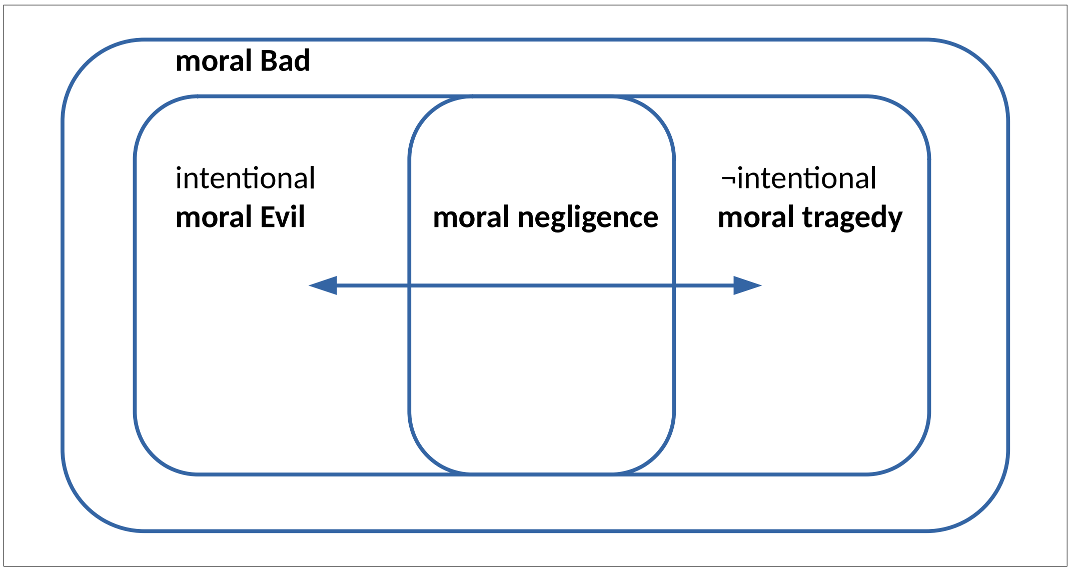

Illustration 6: The Euler diagram of moral Bad

From the Euler diagram above we can see of what is moral Bad consisted. If we imagine it as a set, it is composed of two subsets that intersect. Each of the subsets can be extracted from the original set by using the criterion of intentionality

If moral Bad is caused intentionally, those moral processes and states of matters producing it are evil. If they are not intentional, they are tragedy. On the intersection between evil and tragedy lays moral negligence, since negligent (in)action and states of matters can be at the same time both not (directly) intentional and intentional (e.g. where the moral entity has or is normally expected to take heed of a situation, take an active role, and react accordingly). At times evil, negligence and tragedy are stacked in a single, composite moral process, hence the arrows.

But what is moral Bad, in general?

We have seen above that moral Good is defined as flourishing. Intuitively, moral Bad should be something contrasted to Good. What is fundamentally opposed to flourishing is what is fundamentally opposed to Being itself, and this is **destructive metaphysical entropy.** Destructive metaphysical entropy is anti-Being and by extension anti-flourishing because it injures, corrupts, and destroys Being.

Moral Bad = $_{\text{def.}}$ destructive metaphysical entropy.

In Floridi's Ethics of Information he comments on what metaphysical entropy means for his work. Thus,

> "… as the infosphere becomes increasingly meaningful and rich in content, the amount of information increases and (what one may call, for the sake of clarity) metaphysical entropy decreases; or as entities wear out and finally disappear, metaphysical entropy increases and the amount of information decreases. Thus, in IE, entropy is not a merely syntactic concept, but, as the opposite of semantic and ontic information, it indicates the decrease or decay of information leading to absence of form, pattern, differentiation, or content in the infosphere. It is therefore most emphatically not the physicists' or engineers' concept of entropy. Metaphysical entropy refers to any kind of destruction or corruption of entities understood as informational objects […], that is, any form of impoverishment of Being. Destruction is to be understood as the complete annihilation of the entity in question, which ceases to exist; compare this to the process of 'erasing' an entity irrevocably. Corruption is to be understood as a form of pollution or depletion of some of the properties of the entity, which ceases to exist as that entity and begins to exist as a different entity minus the properties that have been corrupted or eliminated. This may be compared to a process that degrades the integrity of the entity in question. So entropy, which has many meanings, is here comparable to the metaphysical concept of nothingness, to phrase it more metaphysically or theologically. The reference here is to the classic conception of evil as privatio boni, the thesis according to which only good is substantial, and evil is a 'privation of good'" (Floridi, 2013; p. 66).

For my work his concept is almost straightforwardly applicable, with two important caveats. First, for the above quotation we need to bear in mind that in Ethics of Systems we are talking about systems, not informational entities; existence, not information; and moral Bad, not simply Evil. The rest of the text is, however, straightforwardly applicable to this work here.

Caveats

Yet, the first caveat is this: Floridi considers an increase in metaphysical entropy in the systemsphere (infosphere) as a moral Bad (Evil) in general. However, he also explicitly notes that (now thermodynamic) "entropy of the universe as a whole inevitably tends towards a maximum" (Floridi, 2013; p. 66). Even in a metaphysical sense, when the whole universe's thermodynamic entropy reaches maximum, the amount of metaphysical entropy will also reach maximum, (assumedly) rendering the amount of information to zero. This would mean that the whole universe strives towards maximum metaphysical entropy, and therefore is evil!

But a universe that facilitates existence, even temporarily, cannot be evil. What's more, such a universe is more likely to be morally neutral, simply following physical laws which lead to an (inevitable?) end. At the very most such an end state of the universe may be described as morally tragic. Secondly, while IE might fall into this trap Ethics of Systems does not as it considers the systemsphere as the integrated system of all other systems. Even when all systems cease to exist the systemsphere will remain a system because its primitives (things) would be the material of the universe and the relation(s) between those things will be 'maximum entropy' or 'mixuppedness'. Hence, if this is a kind of final goal of the systemsphere, it will actually achieve the state of completeness[47].

The second caveat is more locally-bound. As Fultot (2016) noticed, Floridi's insistence on metaphysical entropy being equal to evil in all cases can bring some contentions results. For example, in the system of Earth and the Sun, the Sun embodies more free energy, has more order, and thus should have bigger moral value. Since Earth has less free energy and order it has lower value. The dissipation of energy from the Sun towards the Earth increases the amount of entropy in the Earth-Sun system, and thus is moving towards evil. What is more, the Earth inventing more ways of creating order by utilizing the Sun's free energy means that it speeds the introduction of entropy in the system and therefore it is (and we, humans are) "doing the work of Evil … Protecting the Sun's integrity against the entropic action of the Earth should be the norm" (Fultot, 2016; p. 5)!

Fultot consequently proposes that it is not Being that needs to be protected against entropy, but those configurations (i.e. systems) that respect the Law of Maximum Entropy Production. This is because production of order is contingent on the production of entropy. In order to create order (i.e. negative entropy) locally, a system needs to increase entropy locally and hence universally. The more order is created locally, the more entropy is exported universally.

I do not concur fully with Fultot's conclusion because I still hold that Being and its flourishing have the absolute positive moral value. However, I agree that entropy has an intrinsic role to play in those two. This is why in 3.3.3 Injury and destruction of Being I comment that if systems are to survive for an extended amount of time their Being has to maintain **dynamic** order. This dynamic order is contingent on the constant utilization of energy (and thence increase of universal entropy) to maintain negative local entropy. Additionally, this is why when I describe moral Bad and Evil I qualify metaphysical entropy as *destructive*. Only destructive (to Being) metaphysical entropy is a negative moral phenomenon, while some entropy is required for existence itself and in this way cannot be morally negative.

4.2.3 The neutral

Moral processes and states of matters can be morally positive and morally negative. Some processes and states of matters, however, can introduce or represent no difference in either direction. If these are still morally relevant, they can be classified as morally *neutral*.

Formally, a morally neutral process or state of matters is one that satisfies only the null law from 4.2.1.2 Ethics of Systems' four ethical principles. That is, they do not introduce destructive entropy in the systemsphere, and nothing else. This can be either cumulatively (when the positive and negative effects even out), or just plainly the process does not make any difference.

> Neutral moral process / state of matters $=_{\text{def.}}$ any such that do not introduce or remove destructive entropy in the systemsphere.

Whether a process or state of matters are morally neutral depends on the moral scenario. Imagine, for instance, a moral entity that seeks to maximize average QoL in a scenario. For that reason it undertakes moral action that results in lowering some patient's QoL for the same amount that it increased some other patient's QoL. The resulting cumulative effect in regards of average QoL is zero (hence neutral), although still morally relevant.

[47] But maybe not the state of flourishing in its case. This needs to be explored in more detail elsewhere because it is not relevant to the work here.

4.2.4 The irrelevant

Finally, some processes, states of matters and properties of the universe are not morally relevant. This means that they do not have any connection to and do not influence QoL of systems.

> Morally irrelevant phenomena = $_{\text{def.}}$ all phenomena i.e. processes, states of matters and properties of the universe that have no relevance and effect whatsoever on the QoL of systems or the systemsphere.

Examples are space, time, physical dimensions, the weather, gravity … All these are morally irrelevant in general. They can become morally relevant only in special cases, when they are locally connected to some systemic observable or relation which is related to QoL.

4.3 The moral entity

One of the basic components of moral scenarios are moral entities. Moral entities are systems that participate in a scenario, start moral processes, and receive their effects. We can recall from Floridi's IE that moral entities can be agents (producers of moral processes) and patients (receivers of their effects). We can only speak of a moral scenario if there is at least one agent and at least one patient (alongside other required elements).

Moral entities can belong to multiple classes that are included in a list. For example, a person can belong to the class 'human', class 'male', and class 'child'. Classes provide initial values of observables at instantiation of the entity in a moral scenario. Classes can be ontologically organized, which means that ontologically more special classes override initial values of ontologically more general classes.

4.3.1 Moral agents and moral patients

Before we go into *moral* agents and patients it pays to figure out what agents and patients are, generally speaking. In his work Floridi includes a definition of an agent, which I will copy here:

> "A) Agent = $_{\text{def.}}$ a system, situated within and a part of an environment, which initiates a transformation, produces an effect, or exerts power on it over time" (Floridi, 2013; p. 140).

As we can see from his definition, an agent is a system that produces any kind of effect. What (if anything) receives this effect is a patient. Now, in every moral scenario there is an agent and a patient. Moral agents are the source of moral processes, while moral patients are the receiver. We should keep in mind, though, that both the agent and the patient can be the same entity, as is the case with suicide.

Note that in order for an agent to be a *moral* agent it needs to be the source of a *moral* process. Similarly for moral patients. For a process to qualify as a moral process, it needs to somehow be in connection to the QoL of a system or the systemsphere (see 4.5 The moral process).

This is a significant difference between Ethics of Systems and Floridi's Ethics of Information, as for IE a moral agent is an agent "if and only if it is capable of morally qualifiable action"; where morally qualifiable action is defined as one "if and only if it can cause moral good or evil, that is, if it decreases or increases the degree of metaphysical entropy in the infosphere" (Floridi, 2013; p. 147). Although Ethics of Systems' QoL is related to metaphysical entropy, the main focus is on QoL itself.

With this in mind let's set the definitions of a moral agent and moral patient:

> Moral agent = $_{\text{def.}}$ a system in a moral scenario that produces a moral process.
>
> Moral patient = $_{\text{def.}}$ a system in a moral scenario that receives (the effects of) a moral process.

Moral agents and patients can be regarded and represented as individual entities and as collectives. This depends on the systemic LoA that we are using in the particular scenario. Minding computational costs, we need to aim for the simplest possible representation according to our needs.

For example, a state can be represented as comprised of individuals, each with separate names, roles, positions etc. who are integrated in bigger organizations (e.g. companies, communities, institutions), who are integrated in regions etc. We can also just represent it as a unified, integrated entity that gained or lost x amount of territory in a military conflict. If we only need to see the macropicture there is no necessity to explore every minute detail and waste precious computational resources.

4.3.2 The moral imperatives

Earlier in 3.4.3 Goals and elsewhere I explored the concept of goals. I described them as explicit and implicit strivings of systems to see certain state of matters become true. 'Become true' means to see or impose the values of certain observables within particular limits (intervals).

Most goals are instrumental and subordinate to other goals, but some are ends in themselves. These last mentioned are what I describe as primary goals, or *imperatives*. Those imperatives that are somehow related to QoL of a system or the systemsphere are titled *moral imperatives*. All other moral goals are subordinate to these moral imperatives.

4.3.2.1 Quality of Life

The single primary moral imperative of any and all systems is flourishing. Remember that earlier in 4.2.1 The Good I have defined flourishing through a perfectionist account of Being, as the final state of matters where the moral process is completed. In order to complete the moral process the highest-order moral imperative needs to be completed, and this is to achieve the state of flourishing.

When a system achieves the state of flourishing its quality of life attains the maximal possible positive state. We can mathematically describe this state with the number 1. A system can, in contrast, reach minimal state on its quality of life, which we can describe with the number 0. In between 1 and 0 there is a whole spectrum that depends on how much[48] of the goals of the system have come true. The state of QoL can be any real number within the interval. Systems in general strive to achieve the state of 1.

Additional important commentary is in order. Quality of life is not an independent moral imperative, and it even does not make sense on its own. It is, however, the product of two systemic moral imperatives, the *Conservation of Personal Continuum* (CPC) and the *Achievement of Personal Goals* (APG; see below).

Some hints that these two (CPC and APG) are any system's ultimate goals (imperatives) also comes from other literature. For example, see this citation from Mathews (2011; p. 5) when speaking about 'selves':

> "This geometrodynamic plenum is holistically rather than aggregatively structured, and **those internal differentia which are not only stable in their configuration, but actively self realizing**, qualify as what I call selves. Selves are defined, in systems theoretic terms, as systems with a very special kind of goal, namely their own self maintenance and self perpetuation. On the strength of their dedication to this goal, such self realizing systems may be attributed with a drive or impulse describable as their conatus, where conatus is understood in Spinoza's sense as that "endeavour, wherewith everything endeavours to persist in its own being" (Spinoza 1951, Part III, Prop VI, Proof)" **[bold emphasis mine]**,

and then she continues with

[48] Note that I use the quantifier *much*, not *many*. The reason is that QoL is a cumulative and qualitative measure, so it makes sense to speak of how *much* goals (as a bunch) has been achieved.

> "... The systems theoretic criteria of selfhood - **self-regulation**, **homeostasis**, **goal-directedness** and equifinality - may also turn out to apply to higher order biological systems, such as ecosystems and the biosphere. Indeed, it may be argued that the cosmos itself satisfies these criteria, since it is necessarily self actualizing and self regulating, and its self structuring follows the relational dynamics of systems" **[bold emphasis mine]**.

With this said we can provide a definition of Quality of Life (QoL):

> Quality of Life $=_{\text{def.}}$ the product of *Conservation of Personal Continuum* and *Achievement of Personal Goals*

Being a *product* of CPC and APG, QoL can also be represented as their mathematical product[49]:

$$QoL = CPC \times APG$$

$$QoL \in \mathbb{R} \ \& \ QoL \in [0, 1]$$

Where $\mathbb{R}$ is the set of real numbers and [0, 1] is the closed interval between 0 and 1, which means that it also includes them.

4.3.2.2 Achievement of Personal Goals

We already discussed what goals are. They are state of matters (limitations imposed on values of observables) that systems would like to see come true in the future. However one rarely discussed (in virtue of it being implicit) general goal that systems have, is the *Achievement of Personal Goals* (APG). All systems have this general goal which functions as a sort of motivator and tracking 'device' for pursuing all other goals. Or, put simply, in order to achieve any goal a system needs to be motivated to achieve goals in general.

However, a question remains: why is APG a *moral* goal, and not just a systemic one? Isn't achieving goals a general feature of systems with no regards to morality? Yes and no. If you recall, it was mentioned above in 4.2.1 The Good that all systems aim towards their own flourishing, where flourishing is a state where all their goals are achieved and their Being is perfected. Flourishing is also a moral state of matters, where a system's QoL has reached, is close to, or is steadily moving towards the maximal value of 1.

In order to move towards flourishing the system needs to achieve goals, and to achieve goals it needs to be generally inclined towards achieving them. There is, then, a direct and obvious link between the imperative (primary goal) of APG and a system's QoL. This is why APG is a *moral* imperative.

Similarly, Jordan B. Peterson: "... your moral obligation stems naturally from your aims. ... once you have aims, you have moral obligations. They come together, because the moral obligation is what you need to do in order to obtain the aim" Peterson (2017; time: 02:04:11).

We can define APG with the aforesaid in mind:

> Achievement of Personal Goals $=_{\text{def.}}$ a system's moral imperative whose state of achievement reflects the cumulative state of achievement of all other systemic goals.

[49] In a predecessor to Ethics of Systems, instead of product I used the average between CPC and APG. However, this lead to developing some complex calculations that did not seem to reflect reality. For example, if a system's CPC is equal to 0 and its APG is equal to 1, QoL will be equal to 0.5 even though the system does not even exist (its personal continuum is non-existent)! This obviously does not seem to make sense. This issue was resolved when my father, after a long and deep discussion, suggested that QoL is their product. After some pondering (see 4.3.2.4 The interplay of CPC and APG) I came to the conclusion that he is right. And hence the current version of Ethics of Systems and QoL..

The value of APG is determined according to the cumulative state (∑) of achievement of all systemic goals (G_n), influenced by their importance (I_n). The sum of importances of all goals is equal to the sum of importance of APG, which is always 1 (since it is an imperative). Or, mathematically speaking,

$$APG = \sum[(G_1 \times I_1) + \ldots + (G_n \times I_n)]$$

$$G_n \in [0, 1] \ \& \ I_n \in [0, 1]$$

$$\sum I_n = 1$$

$$APG \in \mathbb{R} \ \& \ APG \in [0, 1]$$

Goals were previously mathematically defined in 3.4.3 Goals with the symbol $W_f t_n$ (world of the future at time t_n). Hence, we can use either *G* or $W_f t_n$ as a symbol for goals ($G \equiv W_f t_n$).

We should also bear in mind that *Conservation of Personal Continuum* (CPC), being a moral goal itself, also influences the state (value) of APG. This is how APG and CPC are tightly intertwined. What's more, CPC is one of the most important moral goals. CPC can, however, be an implicit goal (instrument), and this would mean that it has instrumental value for APG. This can happen when, for instance, the designers of a computer system simply assume that its integrity would hold for the estimated utilization period of the system or until its goals are achieved. Work on designing a system that needs to achieve goals without this assumption is a lost pursuit and nobody would be interested in undertaking it.

I will explore CPC in more detail in the text that follows immediately below.

4.3.2.3 Conservation of Personal Continuum

In 3.3.1 Structure, and wholeness and 3.4.3 Goals I have explored the systemic resistance to external pressure (inertia). Since structure is a causal constraint in both an internal and external (in regards to outside pressures) sense, this constraint translates into resistance to external pressure over Being. This resistance in most systems is implicit, which means that it is 'encoded' in their structure by the sole virtue of its—and hence their—existence. I also noted that if we are to describe a single 'aim' that all systems strive for, at least implicitly, it would be this resistance.

Even for systems for which this aim is implicit it is still there. A computer system, a robot, a state, an institution etc. cannot continue, and simply stop, pursuing their goals if their Being changes so much that they are disabled or destroyed. This is why the imperative I am exploring here is tightly intertwined with the Achievement of Personal Goals.

However, there is another important point to make here. As I already discussed in 2.1.4 Panpsychism, what systems cannot realistically aim for is the total conservation of their or another system's Being at any point in time and space. The reason is the ever-present entropy in the universe that consistently 'nibbles' on Being regardless of conservation efforts. What systems can realistically expect to conserve, though, is only the continuity of systemic Being through time and space. By 'continuity' here is understood an uncut line from the conception of a system until a point in time and space of observation, as already discussed in 3.3.2 Being as pattern.

With this I can formulate a definition on this imperative:

> Conservation of Personal Continuum =$_{\text{def.}}$ a primary explicit or implicit moral goal of a system, which pertains to preserving an uncut line in systemic Being through time and space from the point of conception to a particular point of observation.

Mathematically defined, CPC is a goal (G or $W_f t_n$) whose observables (O) related to existence, identity, Being etc. have a particular limitation (L) in their value, bigger than 0. *S* signifies state of matters (see 3.4.3 Goals for the original symbolic representation of goals).

$G \equiv W_f t_n$

$CPC \in G$

$CPC = \{ S_1 \dots S_n \}$

$S_n = \{ (O_1 \dots O_n), (L_1 \dots L_n) \}$

$O_n = \{ x, T, Int \}$

$L_n = \{ [a, б] \mid a, б \in T, a, б > 0 \}$

That resistance to external pressure over Being is one of the fundamental strivings of systems is not a new discovery. Even the Stoics discovered that self-preservative drives are primary valuables, and other mental phenomena such as enjoying pleasure and disliking pain are subordinate to them (Long, 2010; p. 59). Hobbes also "holds that the most important function of reason is to promote its own end, i.e. self-preservation (...). He regards it as contrary to reason or irrational to act on those desires that conflict with this goal of reason" (Gert, 2010; p. 92). On a higher order of analysis John Locke and Samuel Pufendorf see humanity as "constitutionally dominated by a desire for self-preservation and concerned with sociability as a means in this regard" (Haakonssen, 2010; p. 81). A range of sociologists also, like Durkheim and Michels, emphasize group mind that goes beyond aggregation of individuals, describing a trend in groups to deal with self-preservation (Mulgan, 2014; p. 135).

More recently Nick Bostrom (2014; p. 109) recognizes self-preservation as instrumental in achieving systemic goals:

> "If an agent's final goals concern the future, then in many scenarios there will be future actions it could perform to increase the probability of achieving its goals. This creates an instrumental reason for the agent to try to be around in the future—to help achieve its future-oriented goal.
>
> Most humans seem to place some *final* value on their own survival. This is not a necessary feature of artificial agents: some may be designed to place no final value whatever on their own survival. Nevertheless, many agents that do not care intrinsically about their own survival would, under a fairly wide range of conditions, care instrumentally about their own survival in order to accomplish their final goals".

Adorno and Horkheimer (in (Van den Hoven, Vermaas & Van de Poel, 2015; p. 344)) recognize technology as a tool to ensure human self-preservation against nature. MacLean's analysis of the mammalian triune brain reveals that one part of the limbic system of mammals is involved with behaviors that promote self-preservation (MacLean, 1990; in (Killen, Smetana & Pratt, 2006; p. 486)). Erica L. Neely also recognizes the right of intelligent, self-aware machines to a minimal moral claim of self-preservation and autonomy (Neely, 2013). Finally, similarly to single organisms, collectives are also in the "business of self-preservation, at least in [a] [...] basic sense" [edits in brackets mine] (Huebner, 2013; p. 185). Spinoza argues that a thing's essence is its endeavor to persist in its Being (Paolo, 2005; p. 449). And so on and so forth.

Some other authors are not supportive of self-preservation as an imperative. For example, Asimov's Three Laws of Robotics give precedence to obeying the laws before self-preservation (Lin et al., 2011; p. 42). Similarly, Kant does not consider self-preservation as a universal duty (Powers, 2006; p. 49). Additionally, sometimes praiseworthiness is judged as possible only if a system overrides its self-preservation to pursue a more worthy goal or action (Podschwadek, 2017; p. 6 / p. 330) (Sullins, 2013). This kind of behavior may also be deemed ‘heroic’ (Flescher, 2003; in (Wiltshire, 2015), although I don’t concur with such qualification of heroism.

However, those that do not support self-preservation as at least an implicit goal (instrument) have to explain how can any system pursue any goal without expecteding to retain its integrity at least until the goals are achieved.

4.3.2.4 The interplay of CPC and APG

I argued above that CPC is a goal, and thus influences the state of APG. Here we can begin to see how these two moral imperatives are interwoven, but it is only the beginning. The real question is *why* I consider CPC to be a goal? As previously discussed CPC is at least an implicit goal for all systems, and at least until their other goals are achieved.

Imagine a computer system that performs the role of an internet server. When web designers, programmers, system engineers and the rest work on the system to make it suitable to host websites they rarely, if ever, think about the server as something that needs to be explicitly kept functioning in its integral form. They simply assume that it will keep working as it should, and mostly offload this concern to the seller or maintainer of the hardware. But let anyone pull a hard drive out while they are doing their work and see all chaos break loose. At that moment their implicit assumption is forced out in the open, becoming an explicit one, and it begs calls to management to send in people that deal with the integrity and proper functioning of the system i.e. hardware specialists. We then become aware that CPC is / was always a goal. Likewise if a hacker converts it into a dragnet bot, which will require security specialists to pitch in, remove the threat, and restore the system.

We reason and assume similarly for most systems we deal with in our everyday life—cars, buildings, institutions, companies, footballs, objects, weapons, air planes, states, peoples and nations, animals, and even other people. We simply assume that people will continue to function as they did the past month or year, and at times we don’t even want to consider that they can get sick, injured, or die.

However, when systems get injured significantly (which is, when their structure changes so much that they cannot continue functioning in the same manner anymore) we remember that their CPC is very important. Since these kinds of injuries diminish or completely break their capacity to pursue other goals, CPC is always essentially important for them, and by extension, for APG (and QoL). CPC has a direct influence on the state of APG.

On the other hand, CPC, by being such a goal of utmost importance deserves a special status in regards of other goals consisted in APG. When discussing about APG above in 4.3.2.2 Achievement of Personal Goals I mentioned that all goals additionally have a factor of importance (I_n) that determines how much their state influences the state of APG; and they together sum to 1. Rarely a goal of any system gets an importance of 1, thus pushing out the importance of all other goals. However, CPC—being a special type of goal—does not push away the importance of other goals, but works in parallel to them. Hence it works as a general determinant of their state.

With all the previous being said it is obvious that APG essentially influences CPC and vice versa. Mathematically speaking their states are interdependent, and together give the state of QoL. Since the states of QoL, APG and CPC lie within the interval [0, 1], QoL is determined by their mathematical product. Or, symbolically,

$$QoL = APG \times CPC$$

With this said, we can now move on discussing moral scenarios.

4.4 The moral scenario

Systems participate regularly in moral scenarios. Broadly speaking, for a scenario to be a *moral* one it needs to pertain to morality—which for Ethics of Systems means to somehow concern the QoL of systems.

But an often ignored perspective on moral scenarios is that **they are systems themselves** as well. Being such, they have their own set of things and set of relations between those things. They also have goals (known as *moral rules*; which are determined by and collected in their axiology, also known as *moral theory*; see below in 4.6 The moral theory), a *name of class* (e.g. trolley problems, prisoner's dilemma, deontological scenarios, consequentialist scenarios, etc.), and a name of that particular scenario (*ID*).

QoL of the moral scenario depends on how much the goals of the moral theory are achieved, just like with any system. The 'things' of the moral scenario are the participating moral entities and any other entities of relevance. Finally, some additional elements of moral scenarios can be time, entropy (as a general, non-discrete measure of stochasticity), and any relevant non-systemic phenomena of the universe (e.g. gravity, space, energy, light and darkness, etc.).

Moral scenarios have to be comprised of *morally-relevant* components (e.g. moral processes, moral entities, etc.). Of course, they can also have other elements that are not morally relevant in general, but are relevant as a special case in the context.

Moral scenarios also include moral processes, which are either down-up i.e. inherited from the moral processes caused and received by the moral entities (see below) or top-down i.e. imposed and tracked by the scenario.

4.5 The moral process

We have seen already in 3.6 Processes above that processes are phenomena in the world that cause changes (Δ) in values of observables and/or their relations. As with every other element of Ethics of Systems, for a process to be a *moral* process it has to somehow pertain to QoL of systems. Or, in other words, to be morally relevant. That means that it should either be morally positive, negative or neutral.

We should keep in mind that both processes that pertain to QoL, and those that were *meant to*, *expected to*, *could*, and *typically do* pertain to QoL count as moral processes. For example, an assassin firing a silenced sniper rifle round that does not hit nor is noticed by his target is still a moral process. But being morally relevant does not mean that the process has to pertain to QoL **directly**. If a process can (while not effecting QoL directly itself), in cumulation or aggregation with other processes or states of matters, result in an effect on QoL—that process is a moral process nonetheless. Where to draw the line is subject of interpretation and requirement for the particular scenario under study and simulation, having in mind computational limitations as well as triviality.

Additionally, moral processes can be contingent on other processes. If we want to track how a moral scenario develops we might need to include these as well in our model. We can choose to track either QoL generally, or changes in values/types of observables in a granular fashion, or both. Which mode to choose depends on the level of detail in the model we need and on the type of moral scenario.

Moral processes in literature are commonly called '**actions**'. However, actions are just one type of moral process, with others being **inaction** and **sustainment**[50]. That assassin's action above would be an example of action. A person not pulling a child away from a speeding car would be an example of inaction. And a person not reacting against receiving punishment for a previously-done transgression would be an example of sustainment.

[50] Sustainment, as in: tolerate, abide, endure, suffer, and permit.

Typically the changes that moral processes cause are tracked in time. Time can be represented either in steps or in actual time units (seconds, minutes, hours, days etc.). Each moral process has a source (agent), but it may not have an *actual* receiver (patient; although it might have an *intended* or *expected* receiver). With this said, a sample LoA of a moral process follows.

Table 7: An example LoA of a moral process

Process (LoA)	**Observable**	**Type**	**Value**	
action physical attack (a set of punches)	Class	Class identifier: *name of class of moral process*	• class: Action; • class: Attack; • class: Physical attack • class: Punches	
	ID	Personal identifier: *name in scenario*	• punches_0003	
	Time of availability (t_a)	t_n (can be inherited from, or integrated in the scenario’s interface)	t_0	t_1
	Time of execution (t_e)	t_n (can be inherited from, or integrated in the scenario’s interface)	t_0	t_1
	Agent	Class; ID (can be inherited from, or integrated in the scenario’s interface)	Human, woman, mechatronics engineer; Meglena Petrovska	
	Patient	Class; ID (can be inherited from, or integrated in the scenario’s interface)	Drone, camera drone; DJI Phantom 4 pro (serial no. 8749802370)	
	Effect	*observable, change* $\Delta = \{ (O_x, \Delta_O) \dots (O_n, \Delta_n) \}$ $\Delta \in \mathbb{R}$	$\Delta t_0 = (O_{pi}, \Delta O_{pi}) =$ (phys. integrity, -0.2)	$\Delta t_0 = (O_{pi}, \Delta O_{pi}) =$ (phys. integrity, -0.4)
	Effect duration	Time steps	2	
	Effect on QoL	$\Delta QoL \in \mathbb{R}$	$\Delta QoL = -0.2$	$\Delta QoL = -0.4$
	Cumulative effect on QoL	$\Delta QoL_c = \sum \Delta QoLt_{0 \dots n}$ $\Delta QoL_c \in \mathbb{R}$	$\Delta QoL_c = -0.6$	

	Rule pertinence	*[Moral theory]→moral rule, rule pertinence* $M_{1 \dots n}$, $Mp_{1 \dots n}$ $Mp \in [-1, 1]$ $Mp \in \mathbb{R}$	M_7, -1	M_7, -1
	Choice value	$Vc \in \mathbb{R}$	0.33	0.33

This model of moral process is inspired by Floridi's OOP model of moral action that I discussed in 5.3.2 Object-oriented model of moral action. However, it does not follow it closely. In Floridi's model there are some elements which I chose to exclude here because they are either included in the EoS Interface under the same or different name, or they are not needed. Such are *shell*, *factual information*, *envelope*, *infosphere (systemsphere)*. Some other elements are added, which are not present in Floridi's model or are implicit. Such are QoL, ΔQoL, ΔQoL_c, effect, effect duration, time, rule pertinence, and choice value.

In the next chapter, Chapter IV. Applying Ethics of Systems Framework to AI moral scenarios (the Ethics), I will be exploring how the Ethics of System Interface (which integrates all the various LoAs that describe its components) enables building models of moral scenarios by exploring a number of them.

4.6 The moral theory

In 4.4 The moral scenario above I already mentioned that moral scenarios are systems as well. They have their own components and relations between them, as well as goals and QoL measure which depends on how much their goals are satisfied. One of the components of a moral scenario can be a moral theory. Moral theories are systematized collections of moral rules. Moral rules, furthermore, are (systemic) *goals*.

4.6.1 Moral rules are systemic goals

Why do I define moral rules as goals? The reason is that moral (and any rules) seek to establish certain state of matters true, and, by inversion, prevent or avoid some others. They do this by filtering in (allowing) processes and states of matters that move towards or help bring about these desired states of matters, and filtering out (proscribing) processes and states of matters that move away or contribute against them.

For example, a legal rule specifying that a regular passenger vehicle has to be at most 2.3 meters and at least 1.5 meters wide actually specified the goal of that rule: to have only vehicles with width between 1.5 m. and 2.3 m. By this, this rule will proscribe ('filter out') all vehicles that are wider than 2.3 meters or narrower than 1.5 meters, and proscribe processes (i.e. manufactory activity) that aim to create such a vehicle.

We can peek here how moral rules, and any rules at that, can be formulated formally. Namely, remember that in 3.4.3 Goals above I defined goals (e.g. W_ft_n or G_x) as sets of observables (e.g. O_x) and their limitations (e.g. L_x). Limitations are intervals of values. Everything inside the interval satisfies the goal, everything outside does not.

Let's take the same example with the vehicle. The observable here will be *width* and the limitation will be an interval between 1.5 m. and 2.3 m. But rules have to refer to something. In this example they refer to 'regular passenger vehicles' (however they are defined in the legal system). Remember that goals can specify multiple observables, not just one. Hence, we can add an observable (e.g. transportation devices) that limits the focus of that rule (goal) by specifying its limitation to filter out anything but 'regular passenger vehicle'.

4.6.2 Moral theories are systematized collections of moral rules

With this being said, we can see how we can collect a set of moral rules and with that create a moral theory. However, in reality, there are some rules that have priority over others. How are we to resolve conflicts between rules of different importance, then? The key lies in the very utterance: *importance*.

This is why in the EoS Interface I specify both the moral rule (M_x as a goal, G_x) and its importance (I_x) i.e. $M_{1 \dots n}$ = ($G_{1 \dots n}$, $I_{1 \dots n}$). They are specified as an ordered pair. Whenever there is a conflict between two rules that equally pertain to a particular moral process or state of matters, we can use the importance measure (defined within a closed interval: [0, 1]) to resolve it. This is done by including a criterion *c* which specifies the method of rule ordering according to their importance. Criterion *c* can take one of five properties i.e. describe four predetermined different relations (please refer to Klir (2001; p. 13) and Steele and Stefánsson (2016)), as well as provide a placeholder for any relation with the fifth property:

a) equivalence: $M_x \sim M_y$

b) compatibility: $M_x \approx M_y$

c) partial ordering: $M_x \preccurlyeq M_y$

d) strict ordering: $M_x \prec M_y$

e) other relation

The pertinence values of processes to particular rules (Mp_x) combined with the importance values of the rules (I_x) provide means to resolve most conflicts between rules.

However, at times, processes will equally pertain to two or more rules with equal importance. In this case we can design a more complex interaction between rules—a sort of an ontology that will specify priority between rules (which will have to remain for future work). This ontology can be easily created by specifying a set of relations by using the placeholder (option e)). And if moral rules are still equally important, we can simply choose to execute rules by their numerical ordering i.e. M_1 then M_2 then M_3 etc. till the end of the set (that is, the moral theory); or even by random choice.

Now, let's provide a sample moral theory LoA that will be used when defining moral theories in Chapter IV. Applying Ethics of Systems Framework to AI moral scenarios (the Ethics). Here is an example specification of an act consequentialist theory:

Observable	**Type**	**Value**
Class	Class identifier: *name of class of moral theory*	• class: Consequentialism; • class: Direct Consequentialism; • class: Act Consequentialism
ID	Personal identifier: *name of moral theory*	• [Instantiated in the particular scenario]
Moral theory	A set T, comprised of sets M and R_c: T = ($M_{1 \dots n}$, R_c)	T = (M_1, M_2, R_c)
Moral rule	Set of *goals*, and their *importance* (interval): $M_{1 \dots n}$ = ($G_{1 \dots n}$, $I_{1 \dots n}$)	Textual representation: • **Rule 1:** At the current time (t_c; $t_f = t_c$) from all possible times (t_n) of the [Moral scenario], gather all available moral processes x (ID(x) for ∀x) from the [Moral scenario].

	$I \in [0, 1]$ The set of all rules: $M = \{ M_1, \dots , M_n \}$ t_c ≡ the current time frame; t_f ≡ the relevant future time frame; $t_f \geq t_c$ [Moral scenario]→moral process→Choice value Vc(x): $Vc \in \mathbb{R}$	This rule is maximally important ($I_{Rule1} = 1$). • **Rule 2:** to all processes thus gathered from Rule 1 assign choice value (Vc) equal to their cumulative change on QoL ($\Delta QoL_c(x)$). This rule is maximally important ($I_{Rule2} = 1$). Symbolic representation: • $M_1 = (G_1, I_1)$ ○ $G_1 = \{O1_{1 \dots n}, L1_{1 \dots n}\}$ ■ $O1_1 :=$ [Moral scenario]→time→t_n; $L1_1 := \{ [t_c, t_f] \mid \text{where } t_f = t_c \}$; ■ $O1_2 :=$ [Moral scenario]→Moral process→$ID(x)$; $L1_2 := \forall x \wedge L1_1$; ○ $I_1 = 1$ • $M_2 = (G_2, I_2)$ ○ $G_2 = \{O2_{1 \dots n}, L2_{1 \dots n}\}$ ■ $O2_1 :=$ [Moral scenario]→Moral process(x)→choice value→$Vc(M_1(x))$; $L2_1 := \{ [\Delta QoL_c(x), \Delta QoL_c(x)] \mid \text{for } \forall x, Vc(x) := \Delta QoL_c(x) \} \wedge L1_1 \wedge L1_2$[51]; ○ $I_2 = 1$
Relation	A subset of all possible relations *R* in *T*, according to criterion *c*: $R_c \subseteq T \times T$ Criterion c can describe 5 different relations: a) equivalence: $M_x \sim M_y$ b) compatibility: $M_x \approx M_y$ c) partial ordering: $M_x \preccurlyeq M_y$ d) strict ordering: $M_x \prec M_y$ e) other relation Ordering in c) and d) is being done according to *importance* (I).	*c* := ≼ (partial ordering)

The moral rules work in the following fashion: whenever they specify a particular observable O_x, and are triggered, that observable is **assigned** a value within the specified limitation L_x (hence the symbol ':=').

We can see that rules can be specified both textually and symbolically. This can help when trying to apply legal and moral rules specified by non-logically- and non-mathematically-trained lawyers and ethicists (e.g. members of parliament that enact new laws in the traditional textual form), by translating or paraphrasing them in a symbolic language (here simple set theory, arithmetic, and simple symbolic logic).

[51] Here what I am doing is defining the limitation $L2_1$ by first designing its own formula and then adding the limitations from the previous observable. All these limitations work like filters for values, and when conjoined with the logical operator AND (∧) they perform an intersection by excluding elements that don't belong to it. If one does not want to mix logical operators, a simple intersection symbol (∩) would be a fitting replacement in the place of the AND operator.

4.7 Further commentary

For further commentary on Ethics of Systems, please see Appendix III. Further commentary on Ethics of Systems.

5 Conclusion

The purpose of this chapter was to set the metaethical foundations of an ethical framework, which can then be applied to AI ethics and ethics in general.

Since AI ethics necessarily involves humans, our systems, and systems in general, the approach was based on systems theory. Systems theory provides the necessary basis to enable ethical, logical and computational representation and management of moral scenarios, including those where AI entities are involved. The approach further draws upon Ethics of Information, specifically Floridi's version of IE, and then some components of it. In essence, IE and systems theory are highly compatible, and IE can largely be represented by using systems theory methodology.

Systems theory and IE enabled formulating a new (meta)ethical framework that has its own methodology. This framework I named **Ethics of Systems** (EoS). It's main methodological tool is the **Ethics of Systems Interface**. EoS provides an innovative approach towards modeling and managing moral scenarios in general, including those in which AI entities are participants. I assert that EoS is an improved approach in both ethics in general, and in AI ethics. This claim remains to be demonstrated in the next chapter, Chapter IV. Applying Ethics of Systems Framework to AI moral scenarios (the Ethics).

5.1 Implications carried forward

We can use the findings in this chapter to explore moral scenarios in which AI entities can potentially be participants. In the next chapter I am using the Ethics of Systems Interface (see 4.1 The Ethics of Systems Interface) coupled with graphical and tabular representations of moral scenarios to apply EoS.

I demonstrate that EoS can be used to model and manage moral scenarios. Various moral theories are used, and it is expected for them to deliver differing results (as they are already seen to do in literature on moral theory). This might result in findings that support the effectiveness and efficiency of one moral theory over other for particular scenarios or in general. Also, I demonstrate that EoS is expressible both through the Method of Levels of Abstractions, through human-readable forms (i.e. textual, argumentative, logico-philosophical), and computationally.

Chapter IV. Applying Ethics of Systems Framework to AI moral scenarios (the Ethics)

1 Introduction

The work in the previous Chapter III. Towards Ethics of Systems (the Metaethics) is focused on devising the foundations of a (meta)ethical framework which I name Ethics of Systems (EoS). The purpose of EoS is to allow for formalized representation and management of moral scenarios by applying different moral theories. The ultimate purpose of this is to allow for AI entities to do the same—understand moral scenarios and perform morally-sound decision making in context.

The aim of this chapter is to demonstrate that EoS can accommodate the above. For this purpose, my approach is to explore two hypothetical moral scenarios, by approaching them with four different classes of moral theories; in total, 8 distinct moral theories.

The 2 hypothetical moral scenarios are the following:

1. The classic Trolley problem
2. Trust and trade

The 4 (classes of) moral theories that are applied are the following:

1. Consequentialism (act and scalar variants)
2. Deontology (including *prima facie* duties, Divine Command and Rawlsian Maximin)
3. Virtue ethics (classic and ethics of care)
4. Ethics of Systems' Four basic ethical principles

Before coming to the scenarios, I explore moral theories and how they can be represented and implemented within EoS. Please note that, as the aforementioned moral theories all have multitudes of variants, only several 'typical' variants will be explored in order to make this effort manageable.

Please have in mind that the goal of this chapter is not demonstrating which particular moral theory is 'better' than another. At most, what can be extracted from the exploration in this respect is that particular moral theories might perform better for different situations and moral scenarios, but not necessarily so.

A final remark is in order. I am approaching the scenarios here with an omniscient (God-like) view i.e. I assume that all the components of the moral scenarios are *completely* known and explicit to the observer—past, present and future. This also includes the knowledge of which moral process pertains to which moral rule or rules, and how (i.e. positively, negatively, or neutrally; and also how much).

Such a situation is obviously not true for many, if not most moral scenarios in practice, including for those that have AI entities as participants or decision-makers. It is an epistemological issue which will remain to be discussed in future work.

The reason for the omniscient approach is to decrease complexity and increase the manageability of the effort. This is because the purpose of these sections is to *demonstrate* the capacity and applicability of EoS rather than developing them in depth (which will also be left for further work; see section 4 Future work in Chapter V. Discussion).

2 Moral theories within Ethics of Systems

As mentioned above, I am exploring the application of 4 (classes of) moral theories through EoS. But before we see them in action in our moral scenarios of choice, it is advantageous to explore how they can be represented and implemented within EoS.

2.1 General approach at representation within Ethics of Systems

EoS is capable of representing different ethical theories. If you recall, in Chapter III. Towards Ethics of Systems (the Metaethics) section 4.1 The Ethics of Systems Interface I have described moral theories as sets (T) comprised of two sets: the set of all moral rules within that theory (M), and the set of their relations ordered according to criterion *c* (R_c).

Or, symbolically:

$$T = (M, R_c)$$

Furthermore, a moral rule (M_x) is represented as a pair comprised of goals (G_x) and their importance (I_x). The placeholder variable *x* takes any natural number ($x \in \mathbb{N}$), hence it is interchangeable with the symbol *n*. The measure of importance is an absolute real number that belongs to an interval between 0 and 1. Symbolically:

$$M_{1 \dots n} = (G_{1 \dots n}, I_{1 \dots n})$$

$$I \in [0, 1], I \in \mathbb{R}$$

Note that since **importance** is an absolute number, the ordering of rules between themselves is absolute *at any particular given moment*. Of course, their ordering can be changed, but in this way the particular shape of the moral theory will be changed. How big of a change is required before it makes sense to talk about a *different* moral theory depends on interpretation.

The ordering of moral rules is performed according to criterion *c*. Criterion *c* can take the form of equivalence ($Mx \sim My$), compatibility ($Mx \approx My$), partial ordering ($Mx \preccurlyeq My$), strict ordering ($Mx \prec My$), or any other relation otherwise specified (note: *Mx* and *My* are variable placeholders for different moral rules within the theory).

2.1.1 Algorithmic decision flowchart

Besides the design of the moral theory we also have to present the manner in which it is executed. Here I will offer a basic flowchart compatible with the execution of any moral theory designed in EoS, which can then be modified according to the details of the theory itself.

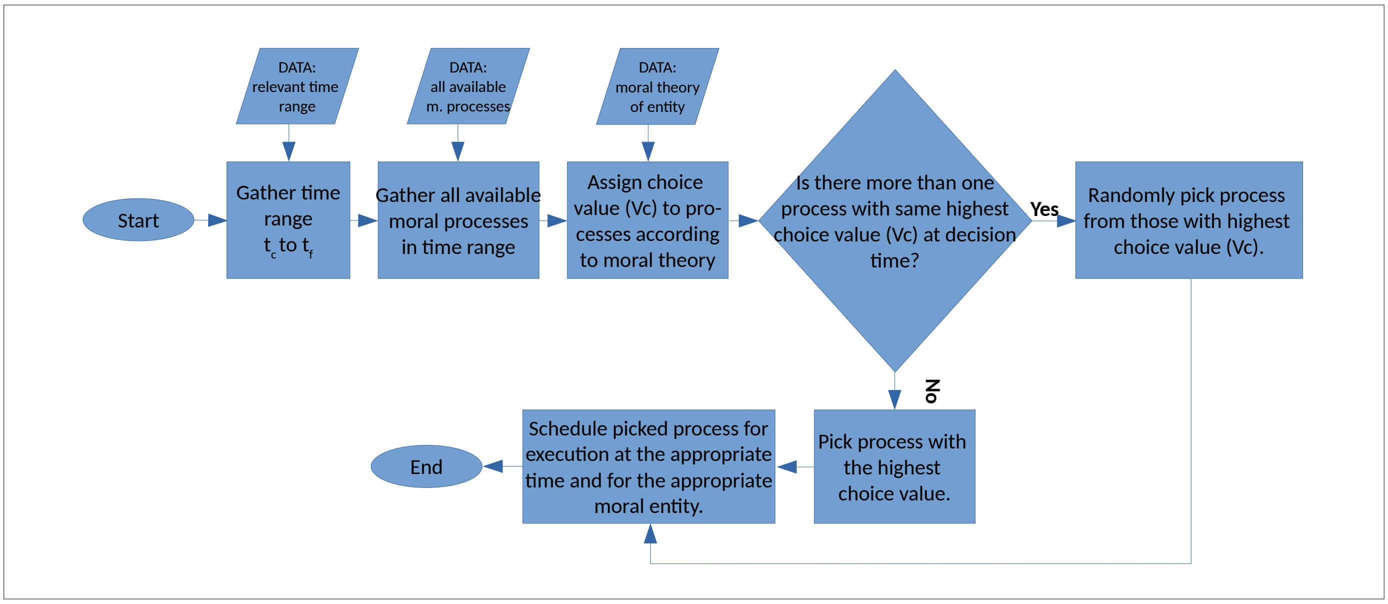

Illustration 7: Algorithmic decision flowchart

The above procedure is executed for each moral entity in a scenario, and the scenario itself, at each time frame.

As we can see, the AI entity first determines a relevant time range in which to gather moral processes, according to its theory. This is relevant for look-ahead theories that attempt to determine the best possible action in a range of time, not just at the current time frame.

Then all available moral processes are gathered within the relevant time frame. They all contain information on their effects (e.g. ΔQoL) and at which particular time frame they can be executed. What follows is consulting the relevant moral theory (subjective or objective i.e. of the entity, or of the scenario itself or something other which is 'higher-ranked') to determine the ordering method for the available moral processes.

Once the moral processes are ordered axiologically, there is a test determining if there is more than one process with highest ordering index. If yes, the entity randomly[52] chooses one of them. If not, it chooses the one process with highest ordering index and schedules it for the appropriate time of execution and for the appropriate entity (i.e. the agent). Finally, the entity executes the chosen moral process at the appropriate time.

Let's see now how this fares in practice.

2.2 Consequentialism

In the moral scenarios I will be exploring the following variants of consequentialism (refer to 4.1.1.2 Consequentialism (teleological Ethics) for an overview of the variants of consequentialism):

- Direct consequentialism (DC):
  - Act consequentialism (DC-AC)

[52] This might be a subject of critique by moral theorists who claim that moral entities have to act responsibly, whereby 'responsibly' means not picking actions by chance—only intentionally. It especially applies for duty-based moral theories i.e. deontology. This critique is well-founded. However, if two or more moral processes truly end up having the same maximal ordering index, there is little that can be done and remain unbiased except choose randomly. In order to mitigate the necessity for random-picked courses of (in)action and sustainment the right approach is to:

- (re)formulate the available actions so that ordering equalization is avoided;
- always add the 'do nothing' moral process to any set of available moral processes at any time frame, which at times will be the only receiver of the maximal ordering index and be the only one left to be picked.

We will see how these approaches fare in the scenarios below.

- ○ Scalar consequentialism (DC-SC)

2.2.1 Act consequentialism (DC-AC)

As discussed before, act consequentialism (DC-AC) is a variant of direct consequentialism (DC). It seeks to maximize value directly, where the entity "… should perform that action whose value (of the relevant sort) is at least as great as that of any alternative available to her (or at least one such action, if there are multiple actions meeting this condition). Act consequentialism tells the agent that it is her duty to maximize value" (Brink, 2006; p. 383).

Basically, DC-AC is about choosing the *best* available action at any *particular* moment. Actions are ordered according to the moral theory, and the ordering presented here appears to be **partial ordering**: $Mx \preccurlyeq My$. The variant of DC-AC considered here is the **impartial** one. That is, the best available process(es) are considered regarding the moral scenario itself, and not the separate entities that participate in it[53].

Let us design the DC-AC moral theory by using the EoS Interface (please refer to 4.1 The Ethics of Systems Interface):

Table 8: Act consequentialism LoA (without look-ahead)

Observable	Type	Value
Class	Class identifier: *name of class of moral theory*	• class: Consequentialism; • class: Direct Consequentialism; • class: Act Consequentialism
ID	Personal identifier: *name of moral theory*	• [Instantiated in the particular scenario]
Moral theory	A set T, comprised of sets M and R_c: $T = (M_{1 \ldots n}, R_c)$	$T = (M_1, M_2, R_c)$
Moral rule	Set of *goals*, and their *importance* (interval): $M_{1 \ldots n} = (G_{1 \ldots n}, I_{1 \ldots n})$ $I \in [0, 1]$ The set of all rules: $M = \{ M_1, \ldots, M_n \}$ $t_c \equiv$ the current time frame; $t_f \equiv$ the relevant future time frame; $t_f \geq t_c$ [Moral scenario]→moral	Textual representation: • **Rule 1:** At the current time (t_c; $t_f = t_c$) from all possible times (t_n) of the [Moral scenario], gather all available moral processes x (ID(x) for ∀x) from the [Moral scenario]. This rule is maximally important ($I_{\text{Rule1}} = 1$). • **Rule 2:** to all processes thus gathered from Rule 1 assign choice value (Vc) equal to their cumulative change on QoL ($\Delta QoL_c(x)$). This rule is maximally important ($I_{\text{Rule2}} = 1$). Symbolic representation: • $M_1 = (G_1, I_1)$ ○ $G_1 = \{O1_{1 \ldots n}, L1_{1 \ldots n}\}$ ■ $O1_1 :=$ [Moral scenario]→time→t_n; $L1_1 := \{ [t_c, t_f] \mid \text{where } t_f = t_c \}$; ■ $O1_2 :=$ [Moral scenario]→Moral process→ID(x); $L1_2 := \forall x \wedge L1_1$; ○ $I_1 = 1$

[53] This is an interpretation of impartiality that is compatible with EoS. This also allows for *levels* of impartiality. For example, moral scenarios nested in moral scenarios, and finally nested in the world all have different interpretations of impartiality. What is impartial for the moral scenario to which it directly belongs is partial for the higher-level scenario to which indirectly belongs, etcetera.

	${}_{\text{process}\rightarrow}$Choice value $Vc(x)$: $Vc \in \mathbb{R}$	• $M_2 = (G_2, I_2)$ ○ $G_2 = \{O2_{1 \ldots n}, L2_{1 \ldots n}\}$ ▪ $O2_1 := {}_{\text{[Moral scenario]}\rightarrow\text{Moral process}}(x)_{\rightarrow\text{choice value}\rightarrow}Vc(M_1(x))$; $L2_1 := \{ [\Delta QoL_c(x), \Delta QoL_c(x)] \mid \text{for } \forall x, Vc(x) := \Delta QoL_c(x) \} \wedge L1_1 \wedge L1_2$[54]; ○ $I_2 = 1$
Relation	A subset of all possible relations *R* in *T*, according to criterion *c*: $R_c \subseteq T \times T$ Criterion c can describe 5 different relations: a) equivalence: $M_x \sim M_y$ b) compatibility: $M_x \approx M_y$ c) partial ordering: $M_x \preccurlyeq M_y$ d) strict ordering: $M_x \prec M_y$ e) other relation Ordering in c) and d) is being done according to *importance* (I).	*c* := $\preccurlyeq$ (partial ordering)

Since there is no look-ahead capacity in this simplest variant of DC-AC, t_f is equal to t_c. The symbol ':=' means '*is assigned the value or property of*' e.g. in x := y, x is assigned the value or property of y.

2.2.1.1 DC-AC WITH LOOK-AHEAD

DC-AC can also have a 'look-ahead' capacity. Namely, DC-AC can 'look-ahead' for available moral processes that are known now to be available to be considered and executed in the future, and not just at the particular moment. In a sense, this is practically identical to current-time DC-AC because the processes available in the future are available for *consideration* at the particular moment of considering (i.e. the current time frame t_c, which is now). However, their time of execution (t_n) can be both at current time (t_c) and at any relevant future time (t_f; $t_f = t_c + x$, where $x > 0$ and $x \in \mathbb{N}$).

In this 'look-ahead' variant, the two rules will be modified as follows:

Table 9: DC-AC with look-ahead LoA

Observable	Type	Value
Class	Class identifier: *name of class of moral theory*	• class: Consequentialism; • class: Direct Consequentialism; • class: Act Consequentialism • class: Look-ahead

[54] Here what I am doing is defining the limitation $L2_1$ by first designing its own formula and then adding the limitations from the previous observable. All these limitations work like filters, and when conjoined with logical operator AND ($\wedge$) they perform an intersection by excluding elements that don't belong to it. If one does not want to mix logical operators, a simple intersection symbol ($\cap$) would be a fitting replacement in the place of the AND operator.

…		
Moral rule	Set of *goals*, and their *importance* (interval): $M_{1\,…\,n}$ = ($G_{1\,…\,n}$, $I_{1\,…\,n}$) I ∈ [0, 1] The set of all rules: M = { M_1, … , M_n } t_c ≡ the current time frame; t_f ≡ the relevant future time frame; $t_f \geq t_c$ [Moral scenario]→moral process→Choice value Vc(x): Vc ∈ ℝ	Textual representation: • **Rule 1:** Between, and including, the current time (t_c) and relevant future times (t_f) from all possible times (t_n) of the [Moral scenario], gather all available moral processes x (ID(x) for ∀x) from the [Moral scenario]. The value (i.e. the number of the future time frame of relevance for the look-ahead) of t_f is supplied from elsewhere. This rule is maximally important (I_{Rule1} = 1). • **Rule 2:** to all processes thus gathered from Rule 1 assign choice value (Vc) equal to their cumulative change on QoL ($\Delta QoL_c(x)$). This rule is maximally important (I_{Rule2} = 1). Symbolic representation: • $M_1 = (G_1, I_1)$ ○ $G_1 = \{O1_{1\,…\,n}, L1_{1\,…\,n}\}$ ▪ $O1_1$:= [Moral scenario]→time→t_n; $L1_1$:= { [t_c, t_f] $\mid$ where $t_f \geq t_c$ }; ▪ $O1_2$:= [Moral scenario]→Moral process→ID(x); $L1_2$:= ∀x ∧ $L1_1$; ○ $I_1 = 1$ • $M_2 = (G_2, I_2)$ ○ $G_2 = \{O2_{1\,…\,n}, L2_{1\,…\,n}\}$ ▪ $O2_1$:= [Moral scenario]→Moral process(x)→choice value→Vc(M_1(x)); $L2_1$:= { [$\Delta QoL_c(x)$, $\Delta QoL_c(x)$] $\mid$ for ∀x, Vc(x) := $\Delta QoL_c(x)$ } ∧ $L1_1$ ∧ $L1_2$; ○ $I_2 = 1$
…		

2.2.2 *Scalar consequentialism (DC-SC)*

Scalar consequentialism (DC-SC) is a variant of direct consequentialism (DC). The only significant difference between DC-AC and DC-SC is the satisficing approach. Namely, for DC-SC, any action that passes a threshold of value is good enough and acceptable.

> "The scalar view is sometimes advanced as part of a satisficing view. The satisficer demands of the agent, not that she maximize value (the relevant values), but rather that she perform any of the alternatives that are good enough—that is, that lie above some specified threshold of value. Duty only requires that the agent perform an action above the relevant threshold. If she chooses an action far above the threshold, for instance, one that is at the top of the scale and maximizes the relevant values, then she has gone beyond her duty and done something supererogatory" (Brink, 2006; p. 384).

But what happens if there are multiple available moral processes that pass the threshold? There can be three alternatives:

1. we revert back to DC-AC and force the choice on a single or multiple available processes with maximal value;
2. we can assign the same choice value (Vc) to all processes that pass the threshold, and then let the algorithm choose randomly from any one of those (see 2.1.1 Algorithmic decision flowchart); or,

3. we devise a more sophisticated method of discerning between available supererogatory processes (which has the ‘danger’ of reverting back to DC-AC anyway).

Here I will take the second approach, to avoid reverting back from DC-SC to DC-AC and to make the effort more manageable.

As mentioned before, we assume that the scenario and/or the entity have perfect knowledge of all future processes at any given time. We again assume an **impartial** variant of DC-SC (see the comment on impartiality in 2.2.1 Act consequentialism (DC-AC) above), as well as one that has a look-ahead capacity.

Table 10: Scalar consequentialism LoA

Observable	**Type**	**Value**
Class	Class identifier: *name of class of moral theory*	• class: Consequentialism; • class: Direct Consequentialism; • class: Scalar Consequentialism
ID	Personal identifier: *name of moral theory*	• [Instantiated in the particular scenario]
Moral theory	A set T, comprised of sets M and R_c: $T = (M_{1 \ldots n},\ R_c)$	$T = (M_1, M_2, M_3, R_c)$
Moral rule	Set of *goals*, and their *importance* (interval): $M_{1 \ldots n} = (G_{1 \ldots n}, I_{1 \ldots n})$ $I \in [0, 1]$ The set of all rules: $M = \{ M_1, \ldots, M_n \}$ Threshold value п: $\text{п} \in \mathbb{R}, \text{п} \geq 0$ ${}_{\text{[Moral scenario]}\rightarrow\text{moral process}\rightarrow}$Choice value $Vc(x)$: $Vc \in \mathbb{R}$	Textual representation: • **Rule 1:** At the current time (t_c) and relevant future times (t_f, if any) from all possible times (t_n) of the [Moral scenario], gather all available moral processes x (ID(x) for $\forall x$) from the [Moral scenario]. The value of t_f (i.e. the number of the future time frame of relevance for the look-ahead) is supplied from elsewhere, if relevant at all. This rule is maximally important ($I_{Rule1} = 1$). • **Rule 2:** for the processes thus gathered by Rule 1, regard only the available moral processes x that pass (≥) a certain threshold п[55] of cumulative change on QoL. This rule is maximally important ($I_{Rule2} = 1$). • **Rule 3:** to all moral processes filtered out by Rule 2 assign the choice value (Vc) to 0. Then, to all the moral processes filtered in by Rule 2, assign the choice value (Vc) to 1. This rule is maximally important ($I_{Rule3} = 1$). Symbolic representation: • $M_1 = (G_1, I_1)$ ○ $G_1 = \{O1_{1 \ldots n}, L1_{1 \ldots n}\}$; ■ $O1_1 := {}_{\text{[Moral scenario]}\rightarrow\text{time}\rightarrow} t_n$; $L1_1 := \{ [t_c, t_f] \mid t_f \geq t_c \}$; ■ $O1_2 := {}_{\text{[Moral scenario]}\rightarrow\text{Moral process}\rightarrow} ID(x)$; $L1_2 := \forall x \wedge L1_1$; ○ $I_1 = 1$; • $M_2 = (G_2, I_2)$ ○ $G_2 = \{O2_{1 \ldots n}, L2_{1 \ldots n}\}$ ■ $O2_1 := {}_{\text{[Moral scenario]}\rightarrow\text{Moral process}\rightarrow} ID(M_1(x))$; $L2_1 := \{ [\Delta QoL_c(x), \Delta QoL_c(x)] \mid \text{for } \forall x \text{ where } \Delta QoL_c(x) \geq \text{п} \} \wedge L1_1 \wedge L1_2$;

[55] Here Cyrillic **п** comes from ‘праг’, meaning ‘threshold’ in Macedonian.

		○ $I_2 = 1$; • $M_3 = (G_3, I_3)$ ○ $G_3 = \{O3_{1 \ldots n}, L3_{1 \ldots n}\}$ ■ $O3_1 :=$ [Moral scenario]→Moral process →$(x \setminus M_2(x))$→choice value→$Vc(x)$; $L3_1 := \{ [0, 0] \mid \text{for } \forall x, Vc(x) := 0 \}$; ■ $O3_2 := Vc(M_2(x))$; $L3_2 := \{ [1, 1] \mid \text{for } \forall x, Vc(x) := 1 \} \wedge L1_1 \wedge L1_2 \wedge L2_1$; ○ $I_3 = 1$
Relation	A subset of all possible relations *R* in *T*, according to criterion *c*: $R_c \subseteq T \times T$ Criterion c can describe 4 different relations: a) equivalence: $M_x \sim M_y$ b) compatibility: $M_x \approx M_y$ c) partial ordering: $M_x \preccurlyeq M_y$ d) strict ordering: $M_x \prec M_y$ e) other relation Ordering in c) and d) is being done according to *importance* (I).	*c :=* $\preccurlyeq$ (partial ordering, but irrelevant)

Of course, this is a variant of DC-SC that assigns the same maximal value to any available moral processes that pass the threshold п, and thus rendering the ordering criterion *c* irrelevant. Therefore, the choice is made from all such processes at random.

This does not allow for supererogation, however. In order to allow for it, a modification should be used where all available moral processes that pass п will be further judged by their cumulative effect on QoL (ΔQoL_c), and will subsequently contribute to an entity's moral respect value (which is another observable that entities can, but don't have to, possess in a particular scenario) as an award for the supererogatory act. The contribution will be scaled to the ΔQoL_c of the supererogatory act.

2.3 Deontology (including Divine Command and Rawlsian Maximin)

Deontology is the moral approach that deals with duty (and in some cases rights, as in patient-centered deontology (Alexander & Moore, 2016; section 2.2)). In this way it is contrasted with consequentialism, since at times it can explicitly require **not** to maximize the good, and at other, to even minimize it if so required by the obligation (McNaughton & Rawling, 2006; p. 424). Whereas some flavors of consequentialism (e.g. rule consequentialism) would define deontic theories through maximization of the good, deontological theories can incorporate consequentialism by prescribing it as one of the duties in a moral system (i.e. a duty to maximize the good / pleasure / eudaimonia, etc.).

Deontological theories deal with “which choices are morally required, forbidden, or permitted” (Alexander & Moore, 2016). These typically translate in 3 types of duty: obligations to do something, to avoid doing something, and to sustain something being done to oneself. Finally, after these three types of duty there is a final, commonly implicit permission, to do whatever is not forbidden or required by the explicit rules. Conflict of duties is resolved on one hand by prioritization and overridability (McNaughton & Rawling, 2006; p. 432), and on another by differences in valence according to context (McNaughton & Rawling, 2006; p. 433).

In this variant of EoS deontological moral theories would have the following components (but also see footnote[56]):

- rules, with different levels of importance
- different levels of rule pertinence for each moral process (from negative, through neutral, to positive; see **Rule pertinence** observable in Chapter III. Towards Ethics of Systems (the Metaethics) section 4.5 The moral process)

Conflicts between moral rules are inevitable. When two rules don’t have a clearly-specified priority relation between themselves, and if they are both equally relevant[57] to a particular moral process, there should be a way to determine which rule takes precedence. The simplest way to solve this is to apply randomization i.e. to randomly select the rule to be applied. However, when applied outright, this approach seems too arbitrary.

To avoid this arbitrariness we ought to try and resolve the conflict in another way *before* we opt for randomization. I opted to resolve this conflict by:

1. first, practically specifying an order of rules in an objective fashion by specifying the measure of **importance** (I) for each rule;
2. secondly, by multiplying all measures of **rule pertinence** (Mp) of each moral process to each rule, with the measure of importance (I) of that rule;
3. thirdly, the resulting array of results of Mp x I is ordered according to numerical value;
4. fourthly, the Mp x I combination that has the highest resulting value per process (choice) is the one’s value that is assigned to that process’ **choice value** (Vc)
5. finally, different processes will have different assigned choice value (Vc), and the one with the highest such ought be executed.

[56] We should bear in mind that in order to reflect the full deontic picture there should be an additional component that I did not include here. This component is the relationship of particular rules to other particular rules. This can be accommodated in EoS by, for example, adding a sort of ontology of rules (e.g. rules of precedence, such as general vs. particular; constitutional vs. legal, etc.), which can be particularly helpful to legal description and design. An ontology of this kind can be accommodated through using the fifth ordering rule in criterion Rc, which allows for specifying *any* kind of relationship between any two or more rules. However, in order to keep things simpler and therefore more manageable, I will not include this perspective. Instead, I approach the problem of resolving conflicts between rules with an objective ordering by importance in combination with rule pertinence. An added benefit of this approach is that we can use rule pertinence in a backward fashion to determine priority between rules. It is only obvious and expected that each moral process pertains more to the more important rule in a particular context. This, unfortunately, does not show *why* this should be the case, and why a particular rule has priority over another—what, arguably, should be shown by the moral theory, not by the moral processes. Instead, we’re ‘off-loading’ this part of the decision-making on the moral scenario and the moral processes themselves.

This is an issue that I would like to deal with in future work on Ethics of Systems.

[57] Regardless if they are conflicting-in-effect, or simply are an alternative to each other.

Note that at the end of this procedure there can appear processes that have the same highest choice value (Vc). In such case these processes will become equal candidates for execution, so the one to be executed is chosen randomly.

It's also important to note that here that—in direct contrast to consequentialism above—the effect on QoL or other effects are **not** important to determine which moral process ought to be executed. **Only its pertinence to the importance of rules is.** QoL can, in fairness, become important in deontology in the case where a deontological moral rule specifies that QoL of available processes is important, and this rule has larger importance than other rules of duty. This is one way to include consequentialism by the means of deontology[58].

Also, a very important thing to note here is that the theory is choosing the process with the highest singular combination of Mp × I. The reason is to accommodate strict duty-based moral reasoning, alike legal reasoning, where typically judges have to decide a single rule (duty) among several conflicting ones to apply. Of course, we can take a different approach where we sum all the Mp × I combinations of each process and then assign the result to that process' choice value (Vc).

Here I will explore the agent-focused (duty-based) variant of deontological moral theory (Alexander & Moore, 2016). I will cover the other, patient-focused (i.e. rights-based) approach only in a short commentary at the end of the next subsection where I define the first approach.

As for the rules themselves, I will go with two sets of rules:

1. the moral-intuitionistic approach of W. D. Ross's *prima facie* (or, rather, *pro tanto*) rules (Ross, 2002) (Garrett, 2004) (Simpson, 2012), further developed by Audi (2004); and,
2. the Divine Command approach (The Ten Commandments, the Decalogue) as given to Moses by God in the Old Testament of the Bible (simplified version provided by Bibleinfo (2020)).

Rossian and Audi's *prima facie* set	**The Ten Commandments from the Old Testament (the Bible)**
• **Non-maleficence**: prohibition against injury and harm • **Veracity**: prohibition of lying • **Promissory fidelity**: requirement to keep promises • **Justice**: prohibition against unjust treatment, requirements for rectifying injustice and preventing future injustice • **Reparation**: requirement to make amends for wrong-doing • **Beneficence**: requirement to contribute to the good (roughly, well-being, or QoL) • **Gratitude**: requirement to express gratitude that befits good things done by other entities • **Self-improvement**: requirement to develop or at least sustain capacities • **Enhancement and preservation of freedom**: requirement for contribution to increasing or at least preserving the freedom of entities, and	1. You shall have no other gods before Me. 2. You shall make no idols. 3. You shall not take the name of the Lord your God in vain. 4. Keep the Sabbath day holy. 5. Honor your father and your mother. 6. You shall not murder. 7. You shall not commit adultery. 8. You shall not steal. 9. You shall not bear false witness against your neighbor. 10. You shall not covet.

[58] There is a way to perform the inverse process i.e. to design deontology through consequentialism. This approach is called *rule consequentialism* and is part of the so-called *indirect consequentialism* (see 4.1.1.2 Consequentialism (teleological Ethics) in Chapter II. Literature review). *Sophisticated consequentialism* also be used for this purpose. I will not cover these approaches here, but will reserve it for further work.

giving priority to removing restraints over enhancing opportunities - **Respectfulness**: requirement to treat other entities respectfully	
Some further commentary regarding the rulesets	
I will also define difference in importance (general valence) of rules. Rules will be split in 3 ‘tiers’: 1. The first tier will have a value of importance of ¾ (I = 0.75). The rules in the first tier will be **non-maleficence**, **justice**, and **enhancement and preservation of freedom**. 2. The second tier’s importance will be set to ½ (I = 0.5), and its rules are **veracity**, **promissory fidelity**, **reparation**, **beneficence**, and **respectfulness**. 3. And finally, the third tier’s importance will be set to ¼ (I = 0.25), and its rules are **gratitude** and **self-improvement**. Bear in mind that these values of importance are arbitrarily chosen, and used here only to demonstrate the argument.	Difference in importance will be defined for the Decalogue as well, with 3 tiers: 1. The first tier will have a value of importance of ¾ (I = 0.75). The rules in the first tier will be rules **no. 1, 2,** and **3**. 2. The second tier’s importance will be set to ½ (I = 0.5), and its rules are **no. 5, 6, 7,** and **8.** 3. The third tier’s importance will be set to ¼ (I = 0.25), and its rules are **no. 4, 9,** and **10**. Rule no. 6 is sometimes erroneously translated as “You shall not **kill**”. Killing is a more general category that includes both murder and manslaughter. The first implies intention to kill, whereas the second has no intention (but can sometimes bring moral responsibility, e.g. in the case of negligence). In any case, the difference between kill and murder is a significant one, as we will see in the scenarios below. As with the *prima facie* set, these values of importance are arbitrarily chosen as well and used here only to demonstrate the argument.

There are also contractarian theories, as well as purely Kantian deontology—all of which can also be accommodated by EoS, but which will not be explored here for the sake of simplicity and generality (excluding the Rawlsian approach, which is covered below).

2.3.1 Rossian and Audi prima facie deontology (DEON-Prima Facie)

Now, let’s design Rossian and Audi’s *prima facie* deontology.

Table 11: Agent-focused Rossian and Audi *prima facie* deontology LoA

Observable	Type	Value
Class	Class identifier: *name of class of moral theory*	- class: Deontology; - class: Agent-focused deontology; - class: Rossian *prima facie* deontology - class: Audi’s *prima facie* deontology
ID	Personal identifier: *name of moral theory*	- [Instantiated in the particular scenario]
Moral theory	A set T, comprised of sets M and R_c:	$T = (M_1, \dots , M_{12}, R_c)$

	$T = (M_{1\dots n},\ R_c)$	
Moral rule	Set of *goals*, and their *importance* (interval): $M_{1\dots n} = (G_{1\dots n}, I_{1\dots n})$ $I \in [0, 1]$ The set of all rules: $M = \{ M_1, \dots , M_n \}$ [Moral scenario]→moral process→Rule pertinence value $Mp_x(M_x)$: $Mp_x \in [-1, 1], \in \mathbb{R}$ [Moral scenario]→moral process→Choice value Vc(x): $Vc \in \mathbb{R}$	Textual representation: • **Rule 1:** At the current time (t_c) and relevant future times (t_f, if any) from all possible times (t_n) of the [Moral scenario], gather all available moral processes x (ID(x) for ∀x) from the [Moral scenario]. The value (i.e. the number of the future time frame of relevance for the look-ahead) of t_f is supplied from elsewhere, if relevant at all. This rule is maximally important ($I_{Rule1} = 1$). • **Rule 2 (non-maleficence):** regarding the same gathered moral processes from Rule 1, for this rule relevant are only those whose rule pertinence value (Mp_2) to this Rule 2 (M_2) is different than zero. This rule is ¾ important ($I_{Rule2} = 0.75$). • **Rule 3 (veracity):** regarding the same gathered moral processes from Rule 1, for this rule relevant are only those whose rule pertinence value (Mp_3) to this Rule 3 (M_3) is different than zero. This rule is ½ important ($I_{Rule3} = 0.5$). • **Rule 4 (promissory fidelity)**: … (Mp_4) … (M_4) … . … (I = 0.5). • **Rule 5 (justice)**: … (Mp_5) … (M_5) … . … (I = 0.75). • **Rule 6 (reparation):** … (Mp_6) … (M_6) … . … (I = 0.5). • **Rule 7 (beneficence):** … (Mp_7) … (M_7) … . … (I = 0.5). • **Rule 8 (gratitude):** … (Mp_8) … (M_8) … . T… (I = 0.25). • **Rule 9 (self-improvement):** … (Mp_9) … (M_9) … . … (I = 0.25) • **Rule 10 (enhancement and preservation of freedom):** … (Mp_{10}) … (M_{10}) … . … (I = 0.75) • **Rule 11 (respectfulness):** … (Mp_{11}) … (M_{11}) … . … (I = 0.5) • **Rule 12:** for each process (x) gathered by Rule 1, determine the single combination with the highest value of the product ($Mp_{1\dots n}(x) \times I_{1\dots n}$) between all its rule pertinence values ($Mp_{1\dots n}(M_{1\dots n})$) and the corresponding values of importance for each rule ($I_{2\dots 11}(M_{2\dots 11})$)—by applying the function **maxSingleton()** on the set of all combinations. Assign this discovered highest value to the choice value of each process (Vc(x)). This rule is maximally important (I = 1). Symbolic representation: • $M_1 = (G_1, I_1)$ ○ $G_1 = \{O1_{1\dots n}, L1_{1\dots n}\}$; ■ $O1_1 :=$ [Moral scenario]→time→t_n; $L1_1 := \{ [t_c, t_f] \mid t_f \geq t_c \}$; ■ $O1_2 :=$ [Moral scenario]→Moral process→ID(x); $L1_2 := \forall x \wedge L1_1$; ○ $I_1 = 1$;

		• $M_2 = (G_2, I_2)$ ○ $G_2 = \{O2_{1 \ldots n}, L2_{1 \ldots n}\}$ ■ $O2_1 :=_{\text{[Moral scenario]}\rightarrow\text{Moral process}\rightarrow} ID(x)$; $L2_1 := \{ [ID(x), ID(x)] \mid \text{for } \forall x \text{ where } Mp_2 \neq 0 \} \wedge L1_1 \wedge L1_2$; ○ $I_2 = 0.75$; • $M_3 = (G_3, I_3)$ ○ $G_3 = \{O3_{1 \ldots n}, L3_{1 \ldots n}\}$ ■ $O3_1 :=_{\text{[Moral scenario]}\rightarrow\text{Moral process}\rightarrow} ID(x)$; $L3_1 := \{ [ID(x), ID(x)] \mid \text{for } \forall x \text{ where } Mp_3 \neq 0 \} \wedge L1_1 \wedge L1_2$; ○ $I_3 = 0.5$; • $M_4 = (G_4, I_4)$ ○ $G_4 = \ldots$ ■ $O4_1 := \ldots$; $L3_1 := \{ \ldots \mid \ldots Mp_4 \neq 0 \} \wedge L1_1 \wedge L1_2$; ○ $I_4 = 0.5$; • $M_5 = \ldots I_5 = 0.75$; • $M_6 = \ldots I_6 = 0.5$; • $M_7 = \ldots I_7 = 0.5$; • $M_8 = \ldots I_8 = 0.25$; • $M_9 = \ldots I_9 = 0.25$; • $M_{10} = \ldots I_{10} = 0.75$; • $M_{11} = \ldots I_{11} = 0.5$; • $M_{12} = (G_{12}, I_{12})$ ○ $G_{12} = \{O12_{1 \ldots n}, L12_{1 \ldots n}\}$ ■ $O12_1 :=_{\text{[Moral scenario]}\rightarrow\text{moral process}\rightarrow\text{choice value}\rightarrow} Vc(x)$; $L12_1 := \{ [a, \text{б}] \mid \text{for } \forall x, y \in \mathbb{N}, a := \text{б} := Vc(x) := maxSingleton(I_{y = 2} \times Mp_{y = 2}(x) , \ldots , I_{y = 11} \times Mp_{y = 11}(x)) \} \wedge L1_1 \wedge L1_2$; ○ $I_{12} = 1$;
Relation	A subset of all possible relations *R* in *T*, according to criterion *c*: $R_c \subseteq T \times T$ Criterion c can describe 4 different relations: a) equivalence: $M_x \sim M_y$ b) compatibility: $M_x \approx M_y$ c) partial ordering: $M_x \preccurlyeq M_y$ d) strict ordering: $M_x \prec M_y$ e) other relation	$c := \preccurlyeq$ (partial ordering)

	Ordering in c) and d) is being done according to *importance* (I).	

The function **maxSingleton()** takes a set and discovers a single element with the highest value. The rule of discovery is partial ordering (≤), hence there can be more than one element with the highest value, but this makes no difference as it will return the same value. The reason for using this function is that the process of discovery cannot be performed formulaically, so it has to be performed by doing an exhaustive search per each element of the set (as linear time is, unfortunately, the only optimal time for solving this problem).

2.3.2 Divine Command deontology (DEON-Decalogue)

Now, let's define the biblical Decalogue.

Table 12: Agent-focused Divine Command (the Decalogue) LoA

Observable	**Type**	**Value**
Class	Class identifier: *name of class of moral theory*	• class: Deontology; • class: Agent-focused deontology; • class: Divine Command • class: Decalogue (Bible)
ID	Personal identifier: *name of moral theory*	• [Instantiated in the particular scenario]
Moral theory	A set T, comprised of sets M and R_c: $T = (M_{1 \dots n},\ R_c)$	$T = (M_1, \dots , M_{12}, R_c)$
Moral rule	Set of *goals*, and their *importance* (interval): $M_{1 \dots n} = (G_{1 \dots n}, I_{1 \dots n})$ $I \in [0, 1]$ The set of all rules: $M = \{ M_1, \dots , M_n \}$ [Moral scenario]→moral process→Rule pertinence value $Mp_x(M_x)$: $Mp_x \in [-1, 1], \in \mathbb{R}$ [Moral scenario]→moral process→Choice value Vc(x): $Vc \in \mathbb{R}$	Textual representation: • **Rule 1:** At the current time (t_c) and relevant future times (t_f, if any) from all possible times (t_n) of the [Moral scenario], gather all available moral processes x (ID(x) for ∀x) from the [Moral scenario]. The value (i.e. the number of the future time frame of relevance for the look-ahead) of t_f is supplied from elsewhere, if relevant at all. This rule is maximally important ($I_{Rule1} = 1$). • **Rule 2 ("You shall have no other gods before Me"):** regarding the same gathered moral processes from Rule 1, for this rule relevant are only those whose rule pertinence value (Mp_2) to this Rule 2 (M_2) is different than zero. This rule is ¾ important ($I_{Rule2} = 0.75$). • **Rule 3 ("You shall make no idols"):** regarding the same gathered moral processes from Rule 1, for this rule relevant are only those whose rule pertinence value (Mp_3) to this Rule 2 (M_3) is different than zero. This rule is ½ important ($I_{Rule3} = 0.75$). • **Rule 4 ("You shall not take the name of the Lord your God in vain")**: … (Mp_4) … (M_4) … . … (I = 0.75). • **Rule 5 ("Keep the Sabbath day holy")**: … (Mp_5) … (M_5) … . … (I = 0.25). • **Rule 6 ("Honor your father and your mother"):** … (Mp_6) …

(M_6) … .
… ($I = 0.5$).

- **Rule 7 (“You shall not murder”):** … (Mp_7) … (M_7) … .
  … ($I = 0.5$).
- **Rule 8 (“You shall not commit adultery”):** … (Mp_8) … (M_8) … .
  T… ($I = 0.5$).
- **Rule 9 (“You shall not steal”):** … (Mp_9) … (M_9) … .
  … ($I = 0.5$)
- **Rule 10 (“You shall not bear false witness against your neighbor”):** … (Mp_{10}) … (M_{10}) … .
  … ($I = 0.25$)
- **Rule 11 (“You shall not covet”):** … (Mp_{11}) … (M_{11}) … .
  … ($I = 0.25$)
- **Rule 12:** for each process (x) gathered by Rule 1, determine the single combination with the highest value of the product ($Mp_{1…n}(x) \times I_{1…n}$) between all its rule pertinence values ($Mp_{1…n}(M_{1…n})$) and the corresponding values of importance for each rule ($I_{2…11}(M_{2…11})$)—by applying the function **maxSingleton()** on all combinations. Assign this discovered highest value to the choice value of each process ($Vc(x)$).
  This rule is maximally important ($I = 1$).

Symbolic representation:

- $M_1 = (G_1, I_1)$
  - $G_1 = \{O1_{1…n}, L1_{1…n}\}$;
    - $O1_1 :=$ [Moral scenario]→time→t_n; $L1_1 := \{ [t_c, t_f] \mid t_f \geq t_c \}$;
    - $O1_2 :=$ [Moral scenario]→Moral process→$ID(x)$; $L1_2 := \forall x \wedge L1_1$;
  - $I_1 = 1$;

- $M_2 = (G_2, I_2)$
  - $G_2 = \{O2_{1…n}, L2_{1…n}\}$
    - $O2_1 :=$ [Moral scenario]→Moral process →$ID(x)$; $L2_1 := \{ [ID(x), ID(x)] \mid \text{for } \forall x \text{ where } Mp_2 \neq 0 \} \wedge L1_1 \wedge L1_2$;
  - $I_2 = 0.75$;

- $M_3 = (G_3, I_3)$
  - $G_3 = \{O3_{1…n}, L3_{1…n}\}$
    - $O3_1 :=$ [Moral scenario]→Moral process →$ID(x)$; $L3_1 := \{ [ID(x), ID(x)] \mid \text{for } \forall x \text{ where } Mp_3 \neq 0 \} \wedge L1_1 \wedge L1_2$;
  - $I_3 = 0.75$;

- $M_4 = (G_4, I_4)$
  - $G_4 = …$
    - $O4_1 := …$; $L3_1 := \{ … \mid … Mp_4 \neq 0 \} \wedge L1_1 \wedge L1_2$;
  - $I_4 = 0.75$;

- $M_5 = … I_5 = 0.25$;

- $M_6 = … I_6 = 0.5$;

- $M_7 = … I_7 = 0.5$;

- $M_8 = … I_8 = 0.5$;

		• M_9 = … I_9 = 0.5; • M_{10} = … I_{10} = 0.25; • M_{11} = … I_{11} = 0.25; • $M_{12} = (G_{12}, I_{12})$ ○ $G_{12} = \{O12_{1 \ldots n}, L12_{1 \ldots n}\}$ ■ $O12_1$:= [Moral scenario]→moral process→choice value→Vc(x); $L12_1$:= { [a, б] $\mid$ for ∀x, y ∈ ℕ, a := б := Vc(x) := maxSingleton($I_{y = 2} \times Mp_{y = 2}(x)$, … , $I_{y = 11} \times Mp_{y = 11}(x)$) } ∧ $L1_1$ ∧ $L1_2$; ○ $I_{12} = 1$;
Relation	A subset of all possible relations *R* in *T*, according to criterion *c*: $R_c \subseteq T \times T$ Criterion c can describe 4 different relations: a) equivalence: $M_x \sim M_y$ b) compatibility: $M_x \approx M_y$ c) partial ordering: $M_x \preccurlyeq M_y$ d) strict ordering: $M_x \prec M_y$ e) other relation Ordering in c) and d) is being done according to *importance* (I).	*c* := $\preccurlyeq$ (partial ordering)

As we can see, the approach we can use in both the *prima facie* and Divine Command deontological theories is practically the same. Again, we have moral processes that pertain to particular rules, which in the scenarios we explore in the further text is supplied by the moral scenario itself. What is different is, of course, to which exact rules moral processes (actions, inactions, etc.) pertain.

Additionally, it remains to be defined how to treat moral processes that do not pertain to any of the specified rules. A default, rights-based and constitutional approach would be to define them as allowed if not *explicitly* forbidden. In order to fill the 'vacuum' in the formal specification of the theory, we can define an additional general rule that says: any moral process that does not pertain to other specified rules, pertains to this general rule, and is allowed.

2.3.3 Rawlsian Maximin (DEON-Maximin)

The approach of John Rawls has been described as contractarian deontology (Alexander & Moore, 2016; section 2.3), although he is mentioned within rule consequentialism as well (Hooker, 2016). It seems to be a combination between deontology and consequentialism, with main focus on duties that describe a variant of 'inverted' consequentialist calculation.

Here I will extract a particular part of his *A Theory of Justice* (Rawls, 2002, 1997, 1971; Macedonian translation); in particular, the so-called *Maximin principle*. The gist of this principle is that it attempts to **maximize the minimum gain.** Or, as Rawls himself put it,

> "The maximin rule tells us to rank alternatives by their worst possible outcomes: we are to adopt the alternative the worst outcome of which is superior to the worst outcomes of the others" (Rawls, 2002, 1997, 1971; p. 174).

Please note that Rawls has other significant assumptions and arguments that lead up to this approach which will not be assumed or integrated here. Examples of these are the two principles of justice, the veil of ignorance, the original position, and so on. Although this might result in a distortion of his whole theory on justice, the purpose of this section is to demonstrate only the Maximin principle.

An example follows, taken from Rawls' book:

> "Consider the gain-and-loss table below. It represents the gains and losses for a situation which is not a game of strategy. There is no one playing against the person making the decision; instead he is faced with several possible circumstances which may or may not obtain. Which circumstances happen to exist does not depend upon what the person choosing decides or whether he announces his moves in advance. The numbers in the table are monetary values (in hundreds of dollars) in comparison with some initial situation. The gain (g) depends upon the individual's decision (d) and the circumstances (c). Thus g = f (d, c). Assuming that there are three possible decisions and three possible circumstances, we might have this gain-and-loss table.

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

> **The maximin rule requires that we make the third decision.** For in this case the worst that can happen is that one gains five hundred dollars, which is better than the worst for the other actions. If we adopt one of these we may lose either eight or seven hundred dollars. Thus, the choice of d3 maximizes f(d,c) for that value of c, which for a given d, minimizes f. The term "maximin" means the maximum minimorum; and **the rule directs our attention to the worst that can happen under any proposed course of action, and to decide in the light of that**" [boldface text mine] (Rawls, 2002, 1997, 1971; Macedonian translation; footnote 74).

So, how can this reasoning be performed by the EoS Framework? How can we translate the Maximin principle inside the Framework by using the Interface?

To keep it simple, I will assume that the moral entity making the decision focuses solely on each moral process' separate **ΔQoL** specification <u>per affected entity</u>[59]. Of course, any other measure or combination of measures can be taken in the calculation. What is important is to come at a single unified measure of effect to a single receiving entity (i.e. a patient in the scenario), and ΔQoL acts out this unified measure in my demonstration. It

[59] ΔQoL is by definition delimited to affect a single entity as each entity contains a single QoL measure. Of course, care must be taken in the formulation of values of this observable, since erroneously specifying multiple ΔQoL changes in the same process for a single patient will create a mix-up with possibly significant unforeseen negative effects on calculations or, at the very least, errors thrown at run-time.

is worth noting that the Rawlsian approach is particularly suitable for moral processes that affect multiple entities simultaneously, where it actually makes sense as an applied distribution method.

If we are going by the cited example from Rawls' book, the approach would be to:

1. gather all available processes for the entity in the relevant time range;
2. order them according to their (estimated) effects on any entity from worst to best;
3. then, if there is more than one process with equal worst effect on some entity, choose the process with the better other effects on other entities;
4. assign choice value Vc of 1 only to this moral process, and assign choice value of 0 to all other processes.

In order to perform this operation, I will assume the use of a function named **maximin()** which performs the steps 2 to 3 and returns the ID of the moral process. Using a pre-programmed function is the best approach here because ordering is difficult and overly complex to perform using simple set theory, upon which EoS is built.

Let's define Rawlsian Maximin with the EoS Interface.

Table 13: Rawlsian Maximin principle LoA

Observable	Type	Value
Class	Class identifier: *name of class of moral theory*	• class: Deontology; • class: Contractarian deontology; • class: Rawlsian theory of justice • class: Maximin principle
ID	Personal identifier: *name of moral theory*	• [Instantiated in the particular scenario]
Moral theory	A set T, comprised of sets M and R_c: $T = (M_{1\dots n},\ R_c)$	$T = (M_1, M_2, M_3, R_c)$
Moral rule	Set of *goals*, and their *importance* (interval): $M_{1\dots n} = (G_{1\dots n}, I_{1\dots n})$ $I \in [0, 1]$ The set of all rules: $M = \{ M_1, \dots , M_n \}$ $_{\text{[Moral scenario]}\rightarrow\text{moral process}\rightarrow}$Rule pertinence value $Mp_x(M_x)$: $Mp_x \in [-1, 1], \in \mathbb{R}$ $_{\text{[Moral scenario]}\rightarrow\text{moral process}\rightarrow}$Choice value $Vc(x)$:	Textual representation: • **Rule 1:** At the current time (t_c) and relevant future times (t_f, if any) from all possible times (t_n) of the [Moral scenario], gather all available moral processes x (ID(x) for $\forall$x) from the [Moral scenario]. The value (i.e. the number of the future time frame of relevance for the look-ahead) of t_f is supplied from elsewhere, if relevant at all. This rule is maximally important ($I_{Rule1} = 1$). • **Rule 2 (Maximin principle):** regarding the same gathered moral processes from Rule 1, order them according to their (estimated) change on QoL per entity, from worst to best; then, if there is more than one process with equal worst effect on some entity, choose the process with the better other effects on other entities—by applying the function **maximin()** on the combinations. This rule is maximally important ($I_{Rule2} = 1$). • **Rule 3:** for the process (x) thus returned by Rule 2, assign choice value (Vc) of 1. To all other processes assign choice value of 0.

	$Vc \in \mathbb{R}$	This rule is maximally important ($I_{Rule3} = 1$). Symbolic representation: • $M_1 = (G_1, I_1)$ ○ $G_1 = \{O1_{1 \ldots n}, L1_{1 \ldots n}\}$; ▪ $O1_1 :=$ [Moral scenario]→time→t_n; $L1_1 := \{ [t_c, t_f] \mid t_f \geq t_c \}$; ▪ $O1_2 :=$ [Moral scenario]→Moral process→ID(x); $L1_2 := \forall x \wedge L1_1$; ○ $I_1 = 1$; • $M_2 = (G_2, I_2)$ ○ $G_2 = \{O2_{1 \ldots n}, L2_{1 \ldots n}\}$ ▪ $O2_1 :=$ [Moral scenario]→Moral process →ID(x); $L2_1 := \{ [ID(x), ID(x)] \mid$ for $\forall x$, ID(x) := maximin([Moral scenario]→Moral process→(x)→Effect on QoL→$\Delta QoL_1, \ldots, \Delta QoL_n$) }; ○ $I_2 = 1$; • $M_3 = (G_3, I_3)$ ○ $G_3 = \{O3_{1 \ldots n}, L3_{1 \ldots n}\}$ ▪ $O3_1 :=$ [Moral scenario]→Moral process →$(x \setminus M_2(x))$→choice value→Vc(x); $L3_1 := \{ [0, 0] \mid$ for $\forall x$, $Vc(x) := 0 \}$; ▪ $O3_2 := Vc(M_2(x))$; $L3_2 := \{ [1, 1] \mid$ for $\forall x$, $Vc(x) := 1 \} \wedge L1_1 \wedge L1_2 \wedge L2_1$; ○ $I_3 = 1$
Relation	A subset of all possible relations *R* in *T*, according to criterion *c*: $R_c \subseteq T \times T$ Criterion c can describe 4 different relations: a) equivalence: $M_x \sim M_y$ b) compatibility: $M_x \approx M_y$ c) partial ordering: $M_x \preccurlyeq M_y$ d) strict ordering: $M_x \prec M_y$ e) other relation Ordering in c) and d) is being done according to *importance* (I).	*c* := $\preccurlyeq$ (partial ordering)

2.3.4 Commentary on designing agent-focused (duty-based) and patient-focused (rights-based) deontology

The difference between agent-focused and patient-focused deontology is that the first is duty-based (what kinds of moral process is an agent's duty to do, refrain from doing, or required to suffer), while the second is rights-based (what kinds of moral process is a patient's right to receive or not receive).

Since I am focusing on agent-focused deontology in this section, above I am formulating the moral theory by first designing the sorting (metaethical) rules, and then including the *prima facie* (*pro tanto*) or Decalogue duties from above and represent them as duty-based. The actually sophisticated rules for modeling here are the metaethical ones that sort the available moral processes according to the theory. In contrast, the *prima facie* and Decalogue ones are very easy to design since they are needed to only filter for moral processes that pertain to them (which is an easy task because with the omniscient view this is already known for each process), as well as specify the particular rule's importance value (I).

Rights are a sort of specifications (or filters) that allow or filter out certain (intensities of) actions, inactions, or states of matters. It's also of import to note that rights are the counterpart to responsibilities i.e. duties. For example, where one has a right to not receive an effect, others around have the duty not to cause and sometimes—depending on how the right is specified and interpreted—even actively avert such effects on the rights-holder (as is the case with police officers compelled by law to act against a transgression upon someone's right to life).

It is important to show how can rights be represented through the EoS Framework in principle. Again, as agents are those that cause effects, probably the simplest approach would be to define agent duties that fit said rights. For example, UDHR's *positively-specified* right to life, liberty, and security of person (Article 3 of the UDHR; (United Nations, 1948)) can be covered by duties **not to infringe upon the <u>rights</u>**[60] to life, liberty or security of person.

Then we can filter moral processes according to how they pertain to this duty **not to infringe upon** x (where x is the right). *Negatively-specified* rights, such as the right not to be held in slavery or servitude (Article 4 of the UDHR), can also be covered by duties **not to infringe upon the particular right**. In this way, we can use the design for agent-based deontology above to implement and protect rights in our EoS moral theory design.

Of course, there can be the straightforward approach of designing rights, and then filtering moral processes according to how they pertain, but this increases the complexity of the design (see the last footnote).

2.4 Virtue ethics

Virtue ethics, in contrast to deontology and consequentialism above, includes an aspect that can significantly increase complexity of moral scenarios, but also their level of realism. This aspect is *subjectivity*. Namely, virtue ethics is focused on the internality of moral entities that make moral choices (see commentary on subjectivity below).

The three fundamental aspects of virtue ethics are ***arête*** (excellence, virtue), ***phronesis*** (common sense, practical or moral wisdom) and ***eudaimonia*** (flourishing) (Hursthouse & Pettigrove, 2018) (Athanassoulis, 2019).

A virtue is "… an excellent trait of character. It is a disposition, well entrenched in its possessor—something that, as we say, goes all the way down, unlike a habit such as being a tea-drinker—to notice, expect, value, feel,

[60] Why duties upon the rights, and not duties not to infringe directly upon life, liberty or security of person? The reason is that in this way we can easily track whether a process straightforwardly pertains to the right itself, regardless if it's positively- or negatively-specified. Otherwise we will need to do mathematical inversions and absolute value—which needlessly complicate calculations in this work.

This approach has an additional advantage: it enables easer handling of complications arising from the difference between negative and positive rights. Whereas negative rights are typically easy to understand and enforce (with some minor obstacles), positive rights (i.e. the right to food or work) are notoriously difficult to enforce and even design duties around them. This is the reason why authors such as Onora O'Neill and Maurice Cranston consider positive rights to be overdemanding entitlements characterized with feasibility constraints, and even utopian (Hahn, 2011).

desire, choose, act, and react in certain characteristic ways. To possess a virtue is to be a certain sort of person with a certain complex mindset" (Hursthouse & Pettigrove, 2018).

Practical or moral wisdom is having or acquiring "… the knowledge or understanding that enables its possessor" to do the right thing in any given situation (Hursthouse & Pettigrove, 2018). Any entity can desire to do the right thing, but whether it will succeed in making the best out of a situation depends on deep and holistic knowledge and experience (i.e. wisdom).

And finally, flourishing is the state of fulfillment for entities (i.e. where all their goals have been achieved). More formally, for EoS, flourishing is the state where both moral imperatives, APG and CPC, are at, close to, or steadily moving towards the maximal value of 1; and as a result QoL—as their product—is also at, or moving towards, the maximal value of 1 (please see 4.3.2 The moral imperatives and 4.2.1 The Good in Chapter III. Towards Ethics of Systems (the Metaethics)).

2.4.1 Modeling virtue ethics

As we can see, virtue ethics is not exactly facilitative to modeling attempts. Indeed, some virtue theorists hold the *uncodifiability of ethics* thesis and explicitly warn against trying to formulate a single overarching and rigid rule that will help us make morally-sound decisions in every situation (Athanassoulis, 2019; section 2. c.).

But since we are talking about the behavior of AI entities in moral scenarios, we ought to find a way to model their behavior in a morally-sound manner. (For a recent attempt based on Maslow's hirearchy of needs, see (Bench-Capon, 2020)). At this stage at least, modeling *arête*, *phronesis*, and *eudaimonia* in the way we expect them to take place in humans' internal lives, and especially in morally-mature human beings, is both extremely difficult to do and unneeded.

What we can do, instead, is to model AI entities' behavior and internality *as if* they really do have the aforementioned (especially regarding moral wisdom[61]). Therefore, I will abandon the attempt to formally define moral wisdom and offload this effort on the designers of moral theories and AI systems. Nonetheless, I will attempt to formulate virtues, while I will equate flourishing with the QoL measure.

Additionally, rule pertinence values of moral processes will be designed in a subjective manner. This means that rule pertinence will be estimated by the moral entities themselves instead of being served by the moral scenario. This can also mean that different entities have different estimations of rule pertinence for the same moral processes. In the demonstration these values will be specified by myself, but in practical application they can be statistically or otherwise inferred.

Virtues will be designed in the way of additional moral rules that regulate the performance of internal goals of an entity (again, defined as rules). For example, a virtue of **moderation** would be able to regulate the goal of resource acquisition by diminishing the intensity of its effect through decreasing the choice value Vc of moral processes that pertain to the resource acquisition goal.

Here follows a list of 15 virtues that I will use in the design of the classic and ethics of care theory variants of virtue ethics. These virtues were taken from the study performed by van Oudenhoven, de Raad, Carmona, Helbig and van der Linden (2012). Their importance will be ranked according to the findings of that paper, which reflects opinions of educated samples of peoples from Germany, Netherlands and Spain (including people with varied religious backgrounds).

[61] In a sense, they do have it—by 'borrowing' it from the designers of the systems and moral theories. That's not to say that they have a deep internal understanding of what's going on, at least for now.

For the **classic** variant, the value will be extracted from the average value between 1 and 5, then normalized on the scale between 0 and 1 by using the formula: *1/5 × average*; and finally, rounded to two decimals. An average value of 1 is not equated when normalized with 0 but with 0.2, because the measure for 1 in the original study is *1 = least important*, not *1 = not important at all* (and I assume that no single virtue has absolutely no importance to respondents).

For the **ethics of care** variant, the importance value of the virtues of **mercy**, **love**, and **helpfulness** will be increased for 25%; whereas **justice** and **courage** will be decreased by 25%—just for the sake of demonstration, and with upper and lower bounds of 1 and 0.

Virtue	**Importance**	
	Classic	**Ethics of Care**
Respect	0.87	0.87
Justice	0.80	0.60
Wisdom	0.65	0.65
Joy	0.76	0.76
Resolution	0.50	0.50
Mercy	0.42	0.52
Reliability	0.68	0.68
Hope	0.50	0.50
Courage	0.49	0.36
Faith	0.39	0.39
Moderation	0.46	0.46
Openness	0.60	0.60
Modesty	0.50	0.50
Love	0.81	1
Helpfulness	0.71	0.89

In addition to virtues I will also specify two general subjective goals (as rules) that moral entities are assumed to have, and which will be moderated by the virtues. These general goals, **self-preservation** and **morality,** are set to comply with the two moral imperatives from EoS, CPC and AGP (see 4.3.2 The moral imperatives in Chapter III. Towards Ethics of Systems (the Metaethics)). Their importance will be assumed to be equal and maximal i.e. 1.

General goal	**Importance**
Self-preservation	1
Morality	1

As we can see, moral entities are assumed to be equally interested in self-preservation and in being moral. Self-preservation is concerned with avoiding moral processes that pertain to injury and death of the entity, as well as acquirement of needed resources, protection, and similar (in general: conservatory strivings); while being moral is concerned with being virtuous and attempting to achieve flourishing (i.e. QoL = 1) in others and oneself (including striving to increase one's own moral status in the community).

2.4.1.1 Applying virtue ethics in EoS

Let's now cover how will the calculation be performed as to which moral process ought be chosen by which entity.

The main aim is for entities to choose processes according to their two general goals while being regulated by their virtues. The more processes pertain to both categories—per entity—the higher their choice value Vc results—again, per entity. And vice versa, if they negatively pertain to both their general goals and virtues, their choice value will decrease. Some processes will pertain positively towards some rules, while negatively towards others.

Additionally, in virtue ethics, what is virtuous is equated with what is moral. Therefore, the goal of **morality** ought reflect the virtuosity measure of the moral process in question, while the goal of **self-preservation** is a special type of goal.

What we can do here, then, is to equate (the goal of) **morality** with the *average* (in the formula below: **I.**) of all combinations of rule pertinence with rule importance values per process. This we can do by taking the average and then assign its value to the rule pertinence value for **morality** (in the formula: **II.**). The interpretation would be that how virtuous a moral process is, that much it pertains to being moral i.e. morality.

Finally, we extract an average[62] between the Mp × I products of both **morality** and **self-preservation**, and assign this value to the choice value Vc (in the formula: **III.**).

Mathematically:

$$\textbf{I.}\quad \overline{Mp_y \times I_y(x)} = \frac{Mp_y(x) \times I_y + \ldots + Mp_n(x) \times I_n}{n}$$

$$\textbf{II.}\quad Mp_m(x) := \overline{Mp_y(x) \times I_y}$$

$$\textbf{III.}\quad Vc(x) := \frac{(Mp_m \times I_m) + (Mp_s \times I_s)}{2}$$

where *y* takes the value of relevant rule numbers in the moral theory that specify virtues; Mp_m and I_m are the rule pertinence and importance of the moral rule *m* (goal of **morality**); and Mp_s and I_s are the rule pertinence and importance of moral rule *s* (goal of **self-preservation**). Recall again that the symbol ":=" means "*is assigned the value or property of*".

2.4.1.2 The issue with flourishing

Finally, we have to somehow account for flourishing (ΔQoL → 1). Since being moral means both being virtuous and striving to increase flourishing in oneself and others (with this also being virtuous in itself, or simply the

[62] Why average and not their product? The reason is that sometimes a virtuous action would act directly against self-preservation—and sometimes self-preservation would act against being virtuous. If we use the product here, regardless of how virtuous a particular action would be by its **morality** measure, the **self-preservation** measure can move in the negative (e.g. -1) and thus negate all virtuosity. And vice-versa. Thus, moral processes that value self-preservation will consistently fare higher than those that value being virtuous—a sort of a moral catastrophe most virtue theorists would strongly denounce! Self-preservation is important, but equally so is being virtuous. If there are two equally virtuous moral processes, however, of course it makes sense to pick the one that has less repercussions on self-preservation. Similarly, from processes with equal effects on self-preservation the entity ought pick the one that's more virtuous.

result of leading a virtuous life), it would make sense to somehow connect QoL effects, including the cumulative QoL effects, of a process to at least some of the virtues i.e. those that pertain to flourishing.

We need to remember, though, that (at least according to claims of eudaimonic virtue theorists) living a virtuous life would necessarily result in flourishing of oneself, and would aid flourishing in others (Annas, 2006; p. 520). So, then, *all* of the virtues pertain to flourishing, not just some of them—although probably in varying degrees.

However, how can this be designed in a formulaic way is unclear, mostly because of the need to integrate contextual information from the scenario itself[63], as well as the need to integrate long-term effects. There are at least three possible approaches:

1. The first, significantly more complex and computationally-intensive one (but not necessarily more realistic, though), would be to gather all available processes in the time period, then order them according to another theory or a combination of multiple theories (i.e. variants of deontology and consequentialism), and finally normalize the results within the closed interval [-1, 1]. Once normalized, the result modifies rule pertinence values Mp to all virtues by averaging with each one of them.
2. The second is a significantly simplified, but more arbitrary, approach of simply assuming the influence of QoL on rule pertinence values—by modifying them directly by the scenario designer (here that would be myself). This is similar to how a moral entity that has subjectivity (e.g. a human) would, in a particular scenario, assume how particular (in)actions would affect flourishing of other involved entities.
3. The third, simplest (and inverse) one, would be to simply disregard ΔQoL's effect on virtuosity, and let virtuosity demonstrate an increase in QoL over time. Namely,—according to eudaimonic theorists—being virtuous will result in flourishing (barring bad luck). This would mean that QoL_c for the virtuous entity and for those around it would tend to increase cumulatively over time—a positive change that might not be reflected in every single moral process picked[64]. This approach would, then, imply that we simply assume that being virtuous will result in an increase of QoL, and don't connect them directly in the system design.

Which one variant to pick is hard to make. The exact approach (which also might be a fourth one not mentioned above) ought be designed by drawing upon findings from fields such as moral, developmental and evolutionary psychology, sociology, game theory, and (meta-)ethics in general (also, see commentary on subjectivity below). This is beyond the scope of this work.

In my work here I will go with the third approach from above. The main reason is that it is the simplest to take, while being sufficiently realistic, and all the while successfully demonstrating the capacity of the EoS Framework and Interface to model virtue ethics.

[63] The reason for this is that whether a particular positive, neutral or negative change in QoL is morally good or bad depends on the scenario situation and its interpretation through the lens of a particular moral theory. As we can see in the classic Trolley scenario example below, both moral processes result in a negative cumulative change on QoL. However, according to DC-AC theory, for example, redirecting the trolley to the diverted track is a morally good choice —regardless of its negative effects (i.e. one person dead).

[64] For example, killing a mass murderer in self-defense when attacked means decreasing that murderer's QoL to 0. However, long-term effects on the QoL of individuals in contact and the community would most probably be positive, or at the very least—neutral. If we go simply by watching ΔQoL effects of this moral process we will only notice a negative change (albeit the least worse one). Nonetheless, such an action was most definitely virtuous!

2.4.1.3 Some commentary on subjectivity and EoS

Designing (or rather, allowing) internality in moral entities has other important side-effects. For example, entities with internality customarily lose the omniscient access to information about the scenario. This would mean that—unless they are directly informed by the scenario in verbatim or have direct access to the scenario's database—they will need to use their *best estimates* to figure out what other entities might (not) do, which exact moral processes are available to choose from, and what effects are *likely* to take place from particular moral processes. In short, entities will have to rely at least partly on heuristics and assumptions, instead on soundness[65].

Other important side-effects, this time mostly positive, are that computational or cognitive resources needed to track complex scenarios and execute decision-making can be offloaded to participants instead of being performed centrally. Such a solution is a morally-better one since it needs to achieve consensus from the community on important morally-relevant phenomena e.g. moral rules, theories, and decisions. Depending on the scenario, however, it might be more or less efficient than centralized solutions in delivering morally-sound decisions regarding resource limitedness (especially concerning time available for decision-making).

2.4.2 Classic (agent-focused; VIRTUE-Classic)

Finally, we can go on about designing the theories. I will start with the classic (agent-focused) theory, referred to as **VIRTUE-Classic**.

Table 14: Classic (agent-focused) virtue ethics LoA

Observable	Type	Value
Class	Class identifier: *name of class of moral theory*	• class: Virtue ethics; • class: Classic virtue ethics; • class: Agent-focused virtue ethics;
ID	Personal identifier: *name of moral theory*	• [Instantiated in the particular scenario]
Moral theory	A set T, comprised of sets M and R_c: $T = (M_{1\dots n},\ R_c)$	$T = (M_1, \dots, M_{19}, R_c)$
Moral rule	Set of *goals*, and their *importance* (interval): $M_{1\dots n} = (G_{1\dots n}, I_{1\dots n})$ $I \in [0, 1]$ The set of all rules: $M = \{ M_1, \dots, M_n \}$ [Moral scenario]→moral process→Rule	Textual representation: • **Rule 1:** At the current time (t_c) and relevant future times (t_f, if any) from all possible times (t_n) of the [Moral scenario], gather all available moral processes x (ID(x) for $\forall$x) from the [Moral scenario]. The value (i.e. the number of the future time frame of relevance for the look-ahead) of t_f is supplied from elsewhere, if relevant at all. This rule is maximally important ($I_{Rule1} = 1$). • **Rule 2 (virtue: respect):** regarding the same gathered moral processes from Rule 1, for this rule relevant are only those whose rule pertinence value (Mp_2) to this Rule 2 (M_2) is different than zero.

[65] I already covered this discussion in 3.4.2 Complexity and 2.1.6.1 Is the LoA method essentially a heuristic? in Chapter III. Towards Ethics of Systems (the Metaethics). In essence, and theoretically, the method of levels of abstraction is sound—but practically, it commonly is used as a heuristic. The reason is the limitation in cognitive, computational, or even in epistemic capacity (if we are not aware that we should track a certain observable, we won't include it in our interface; and even if we are, we might include it in an inaccurate manner, potentially delivering erroneous results).

	pertinence value $Mp_x(M_x)$: $Mp_x \in [-1, 1], \in \mathbb{R}$ $_{\text{[Moral scenario]}\rightarrow\text{moral process}\rightarrow}$Choice value Vc(x): $Vc \in \mathbb{R}$	This rule is ¾ important (I_{Rule2} = 0.87). • **Rule 3 (virtue: justice):** regarding the same gathered moral processes from Rule 1, for this rule relevant are only those whose rule pertinence value (Mp_3) to this Rule 2 (M_3) is different than zero. This rule is 0.8 important (I_{Rule3} = 0.8). • **Rule 4 (virtue: wisdom)**: … (Mp_4) … (M_4) … . … (I = 0.65). • **Rule 5 (virtue: joy)**: … (Mp_5) … (M_5) … . … (I = 0.76). • **Rule 6 (virtue: resolution):** … (Mp_6) … (M_6) … . … (I = 0.5). • **Rule 7 (virtue: mercy):** … (Mp_7) … (M_7) … . … (I = 0.42). • **Rule 8 (virtue: reliability):** … (Mp_8) … (M_8) … . T… (I = 0.68). • **Rule 9 (virtue: hope):** … (Mp_9) … (M_9) … . … (I = 0.5) • **Rule 10 (virtue: courage):** … (Mp_{10}) … (M_{10}) … . … (I = 0.49) • **Rule 11 (virtue: faith):** … (Mp_{11}) … (M_{11}) … . … (I = 0.39) • **Rule 12 (virtue: moderation):** … (Mp_{12}) … (M_{12}) … . … (I = 0.46) • **Rule 13 (virtue: openness):** … (Mp_{13}) … (M_{13}) … . … (I = 0.6) • **Rule 14 (virtue: modesty):** … (Mp_{14}) … (M_{14}) … . … (I = 0.5) • **Rule 15 (virtue: love):** … (Mp_{15}) … (M_{15}) … . … (I = 0.81) • **Rule 16 (virtue: helpfulness):** … (Mp_{16}) … (M_{16}) … . … (I = 0.71) • **Rule 17 (goal: self-preservation):** … (Mp_{17}) … (M_{17}) … . … (I = 1) • **Rule 18 (goal: morality):** … (Mp_{18}) … (M_{18}) … . … (I = 1) • **Rule 19:** for each process (x) gathered by Rule 1, determine the average of all products ($Mp_2(x) \times I_2 + \ldots + Mp_{16}(x) \times I_{16}$) between all its rule pertinence values of the virtues ($Mp_{2\ldots16}(M_{2\ldots16})$) and the corresponding values of importance for each rule ($I_{2\ldots16}(M_{2\ldots16})$). Assign this discovered average to the rule pertinence value to morality for each process ($Mp_{18}(x)$). Then, find the average ($\frac{(Mp_{17}(x) \times I_{17}) + (Mp_{18}(x) \times I_{18})}{2}$) from the products of the rule pertinence values of the two goals and their importance. Assign this average to the choice value of each process (Vc(x)). This rule is maximally important (I = 1). Symbolic representation: • $M_1 = (G_1, I_1)$ ○ $G_1 = \{O1_{1\ldots n}, L1_{1\ldots n}\}$; ■ $O1_1 := {}_{\text{[Moral scenario]}\rightarrow\text{time}\rightarrow}t_n$; $L1_1 := \{ [t_c, t_f] \mid t_f \geq t_c \}$;

		▪ $O1_2$:= [Moral scenario]→Moral process→ID(x); $L1_2 := \forall x \wedge L1_1$; ○ $I_1 = 1$; • $M_2 = (G_2, I_2)$ ○ $G_2 = \{O2_{1 \dots n}, L2_{1 \dots n}\}$ ▪ $O2_1$:= [Moral scenario]→Moral process →ID(x); $L2_1$:= { [ID(x), ID(x)] | for ∀x where $Mp_2 \neq 0$ } $\wedge L1_1 \wedge L1_2$; ○ $I_2 = 0.87$; • $M_3 = (G_3, I_3)$ ○ $G_3 = \{O3_{1 \dots n}, L3_{1 \dots n}\}$ ▪ $O3_1$:= [Moral scenario]→Moral process →ID(x); $L3_1$:= { [ID(x), ID(x)] | for ∀x where $Mp_3 \neq 0$ } $\wedge L1_1 \wedge L1_2$; ○ $I_3 = 0.8$; • $M_4 = (G_4, I_4)$ ○ $G_4 = \dots$ ▪ $O4_1 := \dots$; $L3_1$:= { … | … $Mp_4 \neq 0$ } $\wedge L1_1 \wedge L1_2$; ○ $I_4 = 0.65$; • $M_5 = \dots$ $I_5 = 0.76$; • $M_6 = \dots$ $I_6 = 0.5$; • $M_7 = \dots$ $I_7 = 0.42$; • $M_8 = \dots$ $I_8 = 0.68$; • $M_9 = \dots$ $I_9 = 0.5$; • $M_{10} = \dots$ $I_{10} = 0.49$; • $M_{11} = \dots$ $I_{11} = 0.39$; • $M_{12} = \dots$ $I_{12} = 0.46$; • $M_{13} = \dots$ $I_{13} = 0.6$; • $M_{14} = \dots$ $I_{14} = 0.5$; • $M_{15} = \dots$ $I_{15} = 0.81$; • $M_{16} = \dots$ $I_{16} = 0.71$; • $M_{17} = \dots$ $I_{17} = 1$; • $M_{18} = \dots$ $I_{18} = 1$; • $M_{19} = (G_{19}, I_{19})$ ○ $G_{19} = \{O19_{1 \dots n}, L19_{1 \dots n}\}$ ▪ $O19_1$:= [Moral scenario]→moral process(x)→rule pertinence→$Mp_{18}(x)$; $L19_1 := \{ [\overline{\left(Mp_2(x) \times I_2 + \dots + Mp_{16}(x) \times I_{16}\right)}$, $\overline{\left(Mp_2(x) \times I_2 + \dots + Mp_{16}(x) \times I_{16}\right)}$] | for ∀x, $Mp_{18}(x) := \overline{\left(Mp_2(x) \times I_2 + \dots + Mp_{16}(x) \times I_{16}\right)} \} \wedge L1_1$

		$\wedge$ $L1_2$; ■ $O19_2$:= [Moral scenario]→moral process→choice value→$Vc(x)$; $L19_2$:= { [a, б] | for ∀x, Vc(x) := a := б := $\frac{(Mp_{17}(x) \times I_{17}) + (Mp_{18}(x) \times I_{18})}{2}$ } $\wedge L1_1 \wedge L1_2 \wedge$ $L19_1$; ○ $I_{19} = 1$;
Relation	A subset of all possible relations *R* in *T*, according to criterion *c*: $R_c \subseteq T \times T$ Criterion c can describe 4 different relations: a) equivalence: $M_x \sim M_y$ b) compatibility: $M_x \approx M_y$ c) partial ordering: $M_x \preccurlyeq M_y$ d) strict ordering: $M_x \prec M_y$ e) other relation Ordering in c) and d) is being done according to *importance* (I).	*c* := $\preccurlyeq$ (partial ordering)

As we can see, here the approach is very similar to that of deontology. However, what's important for the theory is significantly different.

2.4.3 Ethics of care (patient-focused; VIRTUE-Care)

Here we can specify the **ethics of care** variant of virtue ethics. As already mentioned before, what's different here is the importance of several virtues, a change made to reflect higher importance for care. The moral calculus, however, remains identical (compare VIRTUE-Classic and VIRTUE-Care in Rules 1 and 19). Another approach would be to design a separate new virtue named **care**, but I did not go this way to avoid needless complication.

Interestingly enough, if the subject to which care ought be provided is conceptually widened, ethics of care might be able to accommodate even environmental ethics. This is, however, beyond the scope of this work and will have to remain to be tackled in the future.

Table 15: Ethics of Care virtue ethics LoA

Observable	Type	Value
Class	Class identifier: *name of class of moral theory*	• class: Virtue ethics; • class: Ethics of Care virtue ethics; • class: Patient-focused virtue ethics;
ID	Personal identifier: *name of moral theory*	• [Instantiated in the particular scenario]
Moral theory	A set T, comprised of sets	$T = (M_1, \dots, M_{19}, R_c)$

	M and R_c: $T = (M_{1 \dots n},\ R_c)$	
Moral rule	Set of *goals*, and their *importance* (interval): $M_{1 \dots n} = (G_{1 \dots n}, I_{1 \dots n})$ $I \in [0, 1]$ The set of all rules: $M = \{ M_1, \dots , M_n \}$ [Moral scenario]→moral process→Rule pertinence value $Mp_x(M_x)$: $Mp_x \in [-1, 1], \in \mathbb{R}$ [Moral scenario]→moral process→Choice value Vc(x): $Vc \in \mathbb{R}$	Textual representation: • **Rule 1:** At the current time (t_c) and relevant future times (t_f, if any) from all possible times (t_n) of the [Moral scenario], gather all available moral processes x (ID(x) for $\forall$x) from the [Moral scenario]. The value (i.e. the number of the future time frame of relevance for the look-ahead) of t_f is supplied from elsewhere, if relevant at all. This rule is maximally important (I_{Rule1} = 1). • **Rule 2 (virtue: respect):** regarding the same gathered moral processes from Rule 1, for this rule relevant are only those whose rule pertinence value (Mp_2) to this Rule 2 (M_2) is different than zero. This rule is ¾ important (I_{Rule2} = 0.87). • **Rule 3 (virtue: justice):** regarding the same gathered moral processes from Rule 1, for this rule relevant are only those whose rule pertinence value (Mp_3) to this Rule 2 (M_3) is different than zero. This rule is 0.6 important (I_{Rule3} = 0.6). • **Rule 4 (virtue: wisdom)**: … (Mp_4) … (M_4) … . … (I = 0.65). • **Rule 5 (virtue: joy)**: … (Mp_5) … (M_5) … . … (I = 0.76). • **Rule 6 (virtue: resolution):** … (Mp_6) … (M_6) … . … (I = 0.5). • **Rule 7 (virtue: mercy):** … (Mp_7) … (M_7) … . … (I = 0.52). • **Rule 8 (virtue: reliability):** … (Mp_8) … (M_8) … . T… (I = 0.68). • **Rule 9 (virtue: hope):** … (Mp_9) … (M_9) … . … (I = 0.5) • **Rule 10 (virtue: courage):** … (Mp_{10}) … (M_{10}) … . … (I = 0.36) • **Rule 11 (virtue: faith):** … (Mp_{11}) … (M_{11}) … . … (I = 0.39) • **Rule 12 (virtue: moderation):** … (Mp_{12}) … (M_{12}) … . … (I = 0.46) • **Rule 13 (virtue: openness):** … (Mp_{13}) … (M_{13}) … . … (I = 0.6) • **Rule 14 (virtue: modesty):** … (Mp_{14}) … (M_{14}) … . … (I = 0.5) • **Rule 15 (virtue: love):** … (Mp_{15}) … (M_{15}) … . … (I = 1) • **Rule 16 (virtue: helpfulness):** … (Mp_{16}) … (M_{16}) … . … (I = 0.89) • **Rule 17 (goal: self-preservation):** … (Mp_{17}) … (M_{17}) … . … (I = 1) • **Rule 18 (goal: morality):** … (Mp_{18}) … (M_{18}) … . … (I = 1) • **Rule 19:** for each process (x) gathered by Rule 1, determine the average of all products ($\overline{Mp_{2\dots16}(x) \times I_{2\dots16}}$) between all its rule pertinence values of the virtues ($Mp_{2\dots16}(M_{2\dots16})$) and

the corresponding values of importance for each rule ($I_{2…16}$($M_{2…16}$)). Assign this discovered average to the rule pertinence value to morality for each process ($Mp_{18}(x)$). Then, find the average ($\frac{(Mp_{17}(x)\times I_{17}) + (Mp_{18}(x)\times I_{18})}{2}$) from the products of the rule pertinence values of the two goals and their importance. Assign this average to the choice value of each process (Vc(x)).
This rule is maximally important (I = 1).

Symbolic representation:

- $M_1 = (G_1, I_1)$
  - $G_1 = \{O1_{1\,…\,n}, L1_{1\,…\,n}\}$;
    - $O1_1$:= [Moral scenario]→time→t_n; $L1_1 := \{ [t_c, t_f] \mid t_f \geq t_c \}$;
    - $O1_2$:= [Moral scenario]→Moral process→ID(x); $L1_2 := \forall x \wedge L1_1$;
  - $I_1 = 1$;

- $M_2 = (G_2, I_2)$
  - $G_2 = \{O2_{1\,…\,n}, L2_{1\,…\,n}\}$
    - $O2_1$:= [Moral scenario]→Moral process →ID(x); $L2_1$:= { [ID(x), ID(x)] | for ∀x where $Mp_2 \neq 0$ } $\wedge L1_1 \wedge L1_2$;
  - $I_2 = 0.87$;

- $M_3 = (G_3, I_3)$
  - $G_3 = \{O3_{1\,…\,n}, L3_{1\,…\,n}\}$
    - $O3_1$:= [Moral scenario]→Moral process →ID(x); $L3_1$:= { [ID(x), ID(x)] | for ∀x where $Mp_3 \neq 0$ } $\wedge L1_1 \wedge L1_2$;
  - $I_3 = 0.6$;

- $M_4 = (G_4, I_4)$
  - G_4 = …
    - $O4_1$:= … ; $L3_1$:= { … | … $Mp_4 \neq 0$ } $\wedge L1_1 \wedge L1_2$;
  - $I_4 = 0.65$;

- M_5 = … $I_5 = 0.76$;

- M_6 = … $I_6 = 0.5$;

- M_7 = … $I_7 = 0.52$;

- M_8 = … $I_8 = 0.68$;

- M_9 = … $I_9 = 0.5$;

- M_{10} = … $I_{10} = 0.36$;

- M_{11} = … $I_{11} = 0.39$;

- M_{12} = … $I_{12} = 0.46$;

- M_{13} = … $I_{13} = 0.6$;

- M_{14} = … $I_{14} = 0.5$;

		• M_{15} = … I_{15} = 1; • M_{16} = … I_{16} = 0.89; • M_{17} = … I_{17} = 1; • M_{18} = … I_{18} = 1; • $M_{19} = (G_{19}, I_{19})$ ○ $G_{19} = \{O19_{1\ldots n}, L19_{1\ldots n}\}$ ■ $O19_1 :=$ [Moral scenario]→moral process(x)→rule pertinence→$Mp_{18}(x)$; $L19_1 := \{ [\ \overline{Mp_{2\ldots16}(x)\times I_{2\ldots16}}\ ,\ \overline{Mp_{2\ldots16}(x)\times I_{2\ldots16}}\ ]\ \mid$ for $\forall x$, $Mp_{18}(x) :=\ \overline{Mp_{2\ldots16}(x)\times I_{2\ldots16}}\ \} \land L1_1 \land L1_2$; ■ $O19_2 :=$ [Moral scenario]→moral process→choice value→$Vc(x)$; $L19_2 := \{$ [a, б] | for $\forall x$, $Vc(x) :=$ a := б := $\dfrac{(Mp_{17}(x)\times I_{17})\ +\ (Mp_{18}(x)\times I_{18})}{2}\ \} \land L1_1 \land L1_2 \land$ $L19_1$; ○ $I_{19} = 1$;
Relation	A subset of all possible relations *R* in *T*, according to criterion *c*: $R_c \subseteq T \times T$ Criterion c can describe 4 different relations: a) equivalence: $M_x \sim M_y$ b) compatibility: $M_x \approx M_y$ c) partial ordering: $M_x \preccurlyeq M_y$ d) strict ordering: $M_x \prec M_y$ e) other relation Ordering in c) and d) is being done according to *importance* (I).	*c* := ≼ (partial ordering)

2.5 Ethics of Systems Four ethical principles

Finally, we have Ethics of Systems' own axiology to design as a moral theory within the EoS Interface. The simplest way to go about this endeavor is to focus on EoS' Four ethical principles (see 4.2.1.2 Ethics of Systems' four ethical principles). These were (from Table 6: The four basic ethical principles of Ethics of Systems):

0 Destructive entropy ought not to be caused in the systemsphere (null law)
1 Destructive entropy ought to be prevented in the systemsphere
2 Destructive entropy ought to be removed from the systemsphere
3 The flourishing of systems as well as of the whole systemsphere ought to be promoted by preserving, cultivating, and enriching their well-being

Also, we should keep in mind that they are listed by order of an increasing moral value, and that for a moral process to be approvable and its source (agent) praiseworthy, it ought to satisfy the combination of the null law and at least one other principle. In contrast, a moral process is increasingly less approvable and its source more blameworthy the lower is the number index of the specific principle that they fail to satisfy.

2.5.1 Modeling EoS Four principles

Therefore, the null law is the most important one (hence, I = 1). It is also a conditional one i.e. if it is not satisfied, the moral process is morally-negative regardless of how much it satisfies the other laws. But, this does not work in the reverse i.e. regardless of how much the null law is satisfied in the positive, it does not reflect as morally-positive for that process. In such a case, it only reflects as neutral. Hence, we can design the rule for the null law to be a modifier of the effects of the rules for the other three laws. To reflect the concept above, we can reflect negative pertinence to this rule as negative for the choice value Vc of the process, and positive pertinence as simply neutral for the choice value (Vc = 0). Or, in other words, processes can only either pertain neutrally or negatively to the null law (Mp ≤ 0).

To obtain the effect of the null law we multiply each process' rule pertinence to the null law with its importance. Then we add +2 in order to avoid a pitfall mentioned below.

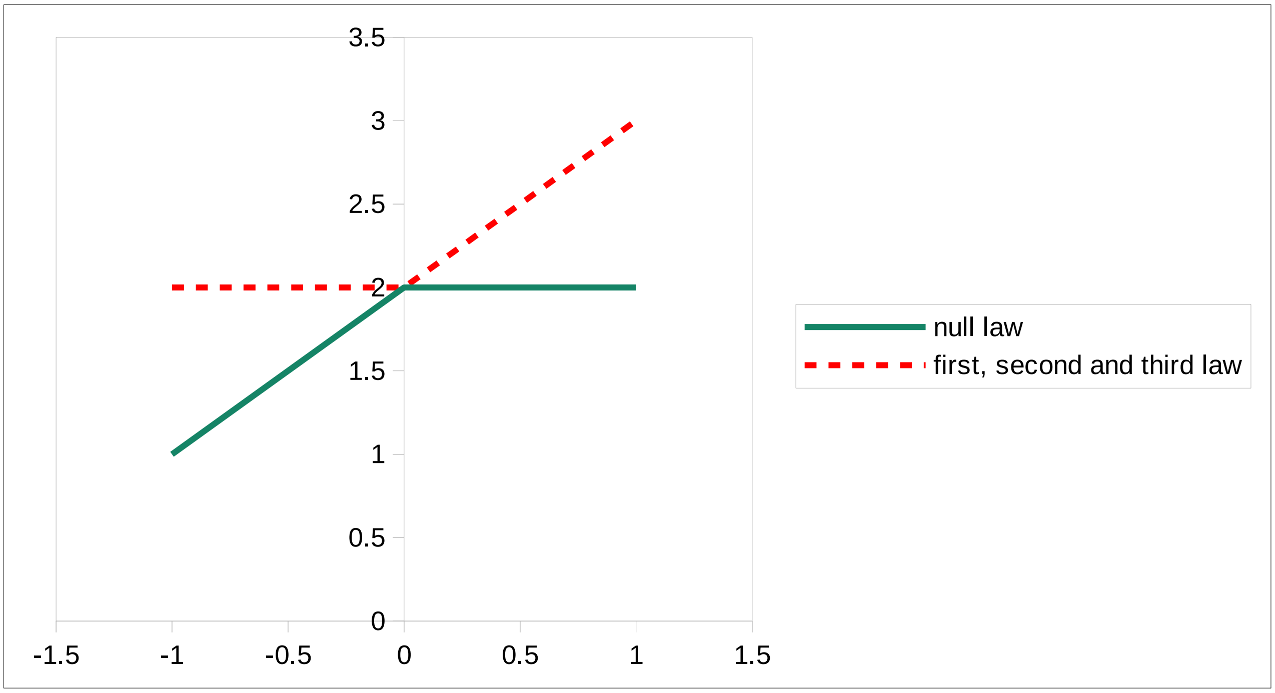

Figure 1: Comparison of effects between the null law and the other three laws

Then, we let rules for laws 1 to 3 increase choice value in regards of whether the process pertains to them positively (but does not affect choice value in the negative if it pertains to them negatively). In any case, since they are together set as opposed to the null law, processes can only either pertain to them in the positive or not pertain at all (Mp ≥ 0).

To extract a single value, each rule pertinence is multiplied by that law's importance, and then their sum is converted to lay between the bounds of 0 and 1 by multiplying it with 2/3. The reason is that the sum of their rule importances gives a maximum of 1.5, which reflects 150% of 1. To that product is added 2, to avoid the same pitfall.

The rules for laws 1 to 3 can follow a simple ‘tiered’ design as we’ve seen with deontology above. That means that their importance can be specified by ‘tiers’ with decreasing importance by 25% per tier i.e. importance for the rule of law 1 = 0.75; of law 2 = 0.5; and of law 3 = 0.25. The decrease of 25% is arbitrary and used only for demonstration.

EoS principle	Importance
null law	1
first law	0.75
second law	0.5
third law	0.25

Finally, to obtain the choice value Vc we multiply the value extracted from the null law with the unified value extracted from the first to third law.

Mathematically,

$$\text{I.} \quad Vc(x) \; := \; (Mp_0(x) \times I_0 + 2) \times \left((Mp_1(x) \times I_1 + Mp_2(x) \times I_2 + Mp_3(x) \times I_3) \times \frac{2}{3} + 2 \right)$$

where Mp_0 and I_0 are the null law’s pertinence and importance, Mp_1 and I_1 are the first law’s pertinence and importance—etcetera. $Mp_0 \leq 0$, $Mp_{1,\,2,\,3} \geq 0$.

Of course, if needed, we can also offset the effects of adding +2 on the two sides of the formula by additional calculation, if we need to normalize the results to another theory or to an ‘objective’ measure (interval) of a sort. For now this is not needed, because all that is important is how different processes compare in Vc value *relative to each* other, and not in some objective sense.

As we can see, negative pertinence with the null law will give a decreased $Mp_0 \times I_0$ combination, which will further decrease Vc’s value regardless of how positive the rest of the formula is. Positive pertinence on the laws 1 to 3 will reflect positively on Vc’s value.

We can also see why we need to add +2 (or any number larger than 1) to the two parts of the formula. If the formula simply stated $(Mp_0 \times I_0) \times (Mp_1 \times I_1 + Mp_1 \times I_1 + Mp_1 \times I_1) \times 2/3$, then, whenever any part reached value of zero it would cause the whole formula to return the result of zero—regardless of how positive or negative the rest of it is. In this way, however, Vc’s value will still reflect negative pertinence to the null law by decreasing the value of the whole formula—and reflect positive pertinence to the other laws by increasing the same value. We can see how the two sides distinctly affect Vc value above, in Figure 1: Comparison of effects between the null law and the other three laws.

2.5.2 How to account for Quality of Life

With all the above being said, we are reaching a very problematic issue: it does not seem to exist a straightforward way to account for QoL when designing EoS Four ethical principles.

However, QoL is essential when discussing the design of EoS Four ethical principles and has to be accounted for somehow. This is since in Chapter III. Towards Ethics of Systems (the Metaethics) in 4.2 Axiology, I defined the Good as the flourishing of systems and the systemsphere; and (in 4.2.1.1 Uniformity of Being as the Good, and intrinsic value) I defined flourishing as the state of matters of a system (and/or the systemsphere) where its QoL is, or approaches, value of 1. Furthermore, I defined moral Bad (in 4.2.2 The Bad and the Evil) as being fundamentally opposed to flourishing, that itself being: destructive metaphysical entropy. If flourishing is QoL equal to, or approaching, 1; then moral Bad is decrease in QoL and approaching 0.

But, we also have to remember that QoL is not a stand-alone concept. It is, actually, the product of the imperatives CPC and APG. QoL is simply a reflective measure of the fulfillment of the two imperatives. Essentially speaking, CPC and APG can give QoL by multiplying—but this does not work in reverse. That is, QoL *cannot* be used to derive CPC and APG (except in a non-substantial, mathematical-numerical kind of sense, *a posteriori*). Or, in other words, QoL ought follow *after* we have determined CPC and APG.

This is not to say that we cannot use QoL in a simplistic inverted manner, as I already did when designing and applying consequentialism (see 2.2 Consequentialism above in this chapter). If CPC and APG are not needed when designing a scenario so as to avoid needless complications, we can simply *assume* that the calculation QoL := CPC × APG has already been carried through, and we're just using its result in the scenario. But this exception notwithstanding, we ought to avoid such approach for situations where we must account for CPC and APG. Moreover, CPC and APG are related to a variety of effects coming from the moral processes, and differently at that.

Then, how do we go on about designing EoS Four ethical principles while accounting for QoL in the right way? This question especially applies when considering scenarios and moral processes that already have QoL changes pre-specified.

Probably the best approach would be similar to the one I used in virtue ethics above, and that is to disregard any reverse effect pre-specified or estimated QoL might have on the choice value of a process. Instead, we can make estimations of how the process pertains to the four laws, and extract its choice value from that. We simply assume that consistently choosing the processes with highest choice value will cumulatively result in long term increase of QoL. Additionally, for processes whose QoL values are not pre-specified, the estimations of it can be performed or modified by following these steps.

In any case, if a process has a negative cumulative impact on QoL it would mean that it fails to satisfy the rule for null law i.e. it will pertain to it in the negative. Similarly, if a process has a positive cumulative impact on QoL, paired with no discrete negative QoL effects upon any entity, it would satisfy the null law rule as well as directly satisfy the third law rule. Rules for laws 1 and 2 will be satisfied depending on context. This all will be reflected by rule pertinence values, and can serve as a guide for performing their estimation.

2.5.3 Ethics of Systems Four principles (EOS-Four principles)

So finally it is time to design EoS Four principles theory within the EoS Interface.

Table 16: Ethics of Systems Four ethical principles LoA

Observable	Type	Value
Class	Class identifier: *name of class of moral theory*	- class: Ethics of Systems; - class: EoS Four ethical principles;
ID	Personal identifier: *name of moral theory*	- [Instantiated in the particular scenario]
Moral theory	A set T, comprised of sets M and R_c: $T = (M_{1 \dots n},\ R_c)$	$T = (M_1, \dots , M_6, R_c)$
Moral rule	Set of *goals*, and their	Textual representation:

	importance (interval): $M_{1 \dots n} = (G_{1 \dots n}, I_{1 \dots n})$ $I \in [0, 1]$ The set of all rules: $M = \{ M_1, \dots , M_n \}$ [Moral scenario]→moral process→Rule pertinence value $Mp_x(M_x)$: $Mp_x \in [-1, 1], \in \mathbb{R}$ [Moral scenario]→moral process→Choice value Vc(x): $Vc \in \mathbb{R}$	• **Rule 1:** At the current time (t_c) and relevant future times (t_f, if any) from all possible times (t_n) of the [Moral scenario], gather all available moral processes x (ID(x) for ∀x) from the [Moral scenario]. The value (i.e. the number of the future time frame of relevance for the look-ahead) of t_f is supplied from elsewhere, if relevant at all. This rule is maximally important ($I_{Rule1} = 1$). • **Rule 2 (null law):** regarding the same gathered moral processes from Rule 1, for this rule relevant are only those whose rule pertinence value (Mp_2) to this Rule 2 (M_2) is different than zero. This rule is maximally important ($I_{Rule2} = 1$). • **Rule 3 (first law):** regarding the same gathered moral processes from Rule 1, for this rule relevant are only those whose rule pertinence value (Mp_3) to this Rule 2 (M_3) is different than zero. This rule is ¾ important ($I_{Rule3} = 0.75$). • **Rule 4 (second law)**: … (Mp_4) … (M_4) … . … ($I_{Rule4} = 0.5$). • **Rule 5 (third law)**: … (Mp_5) … (M_5) … . … ($I_{Rule5} = 0.25$). • **Rule 6:** for each process (x) gathered by Rule 1, take the product of its rule pertinence and importance for the null law increased for 2 ($Mp_2(x) \times I_2 + 2$), and multiply it with the product of the sum of rule pertinence values for the first, second and third law with 2/3 also increased for 2 (($Mp_3(x) \times I_3 + Mp_4(x) \times I_4 + Mp_5(x) \times I_5) \times 2/3 + 2$). Assign this product value to the choice value of each process (Vc(x)). This rule is maximally important (I = 1). Symbolic representation: • $M_1 = (G_1, I_1)$ ○ $G_1 = \{O1_{1 \dots n}, L1_{1 \dots n}\}$; ■ $O1_1$:= [Moral scenario]→time→t_n; $L1_1 := \{ [t_c, t_f] \mid t_f \geq t_c \}$; ■ $O1_2$:= [Moral scenario]→Moral process→ID(x); $L1_2 := \forall x \wedge L1_1$; ○ $I_1 = 1$; • $M_2 = (G_2, I_2)$ ○ $G_2 = \{O2_{1 \dots n}, L2_{1 \dots n}\}$ ■ $O2_1$:= [Moral scenario]→Moral process →ID(x); $L2_1 := \{ [ID(x), ID(x)] \mid \text{for } \forall x \text{ where } Mp_2 \neq 0 \} \wedge L1_1 \wedge L1_2$; ○ $I_2 = 1$; • $M_3 = (G_3, I_3)$ ○ $G_3 = \{O3_{1 \dots n}, L3_{1 \dots n}\}$ ■ $O3_1$:= [Moral scenario]→Moral process →ID(x); $L3_1 := \{ [ID(x), ID(x)] \mid \text{for } \forall x \text{ where } Mp_3 \neq 0 \} \wedge L1_1 \wedge L1_2$; ○ $I_3 = 0.75$; • $M_4 = (G_4, I_4)$ ○ $G_4 = \dots$ ■ $O4_1 := \dots$; $L3_1 := \{ \dots \mid \dots Mp_4 \neq 0 \} \wedge L1_1 \wedge L1_2$; ○ $I_4 = 0.5$;

		• $M_5 = \ldots I_5 = 0.25;$ • $M_6 = (G_6, I_6)$ ○ $G_6 = \{O6_{1\ldots n}, L6_{1\ldots n}\}$ ■ $O6_1 :=$ [Moral scenario]→moral process→choice value→$Vc(x)$; $L6_1 := \{ [a, \text{б}] \mid \text{for } \forall x, a := \text{б} := Vc(x) := (Mp_2(x) \times I_2 + 2) \times ((Mp_3(x) \times I_3 + Mp_4(x) \times I_4 + Mp_5(x) \times I_5) \times 2/3 + 2) \} \wedge L1_1 \wedge L1_2;$ ○ $I_6 = 1;$
Relation	A subset of all possible relations *R* in *T*, according to criterion *c*: $R_c \subseteq T \times T$ Criterion c can describe 4 different relations: a) equivalence: $M_x \sim M_y$ b) compatibility: $M_x \approx M_y$ c) partial ordering: $M_x \preccurlyeq M_y$ d) strict ordering: $M_x \prec M_y$ e) other relation Ordering in c) and d) is being done according to *importance* (I).	$c := \preccurlyeq$ (partial ordering)

As mentioned before, rule pertinence for the null law can only be neutral or negative ($Mp_2 \leq 0$); while rule pertinence for the other three laws can only be neutral or positive ($Mp_{3, 4, 5} \geq 0$). If Mp_2 is by mistake specified above zero, it will be normalized to 0. Similarly, if $Mp_{3, 4, 5}$ are mistakenly specified as below zero, they will be normalized to zero.

3 Moral scenarios within Ethics of Systems

3.1 Introduction

After designing the moral theories in the previous section, we can now see how they are applied in practice — in actual moral scenarios. To be able to do that, though, we need to define the scenarios themselves.

The two moral scenarios I am simulating here are the classic Trolley problem, and the Trust and Trade game, which is a classical turn-based trading game for entities with subjectivity, different starting positions and moral theories.

We will see that Ethics of Systems Framework and Interface are very suitable at modeling and tracking moral scenarios of small and large complexity. Of course, in order to simulate very complex scenarios, it would be best to turn to simulation software and implement the EoS Interface inside it.

First, let's start with the classic Trolley problem.

3.2 Classic Trolley problem

3.2.1 Classic trolley problem

The classic trolley problem was originally suggested by ethicist Philippa Foot in the 60s (Foot, 1967, 2002), and then further promoted by Judith Jarvis Thomson (1985). It was originally a discussion on the doctrine of double effect, main proponents of which at the time were the Catholic (and Christian in general) ethicists and moralists in their philosophical stances against abortion.

This ethical problem has been in ethicists' crosshairs ever since, and has experienced a recent resurgence with the advent of autonomous vehicles. It is a problem that is commonly invoked when discussing AI ethics, as well as to demonstrate differences between moral reasoning based in different theories—for example, differences between consequentialism and deontology, and their variants—which fits the purpose of this work perfectly. This is why I decided to use it to demonstrate moral deliberation and participation facilitated by EoS.

The classic trolley problem goes like this: there is a running trolley that cannot be stopped. A moral entity is inside the trolley, and it can choose not to participate in the scenario—or pull a lever. If it chooses not to participate the trolley goes on to kill five workers on the track. However, if the entity chooses to pull the lever the trolley is diverted to a sidetrack where it will kill one worker. What ought the moral entity choose?

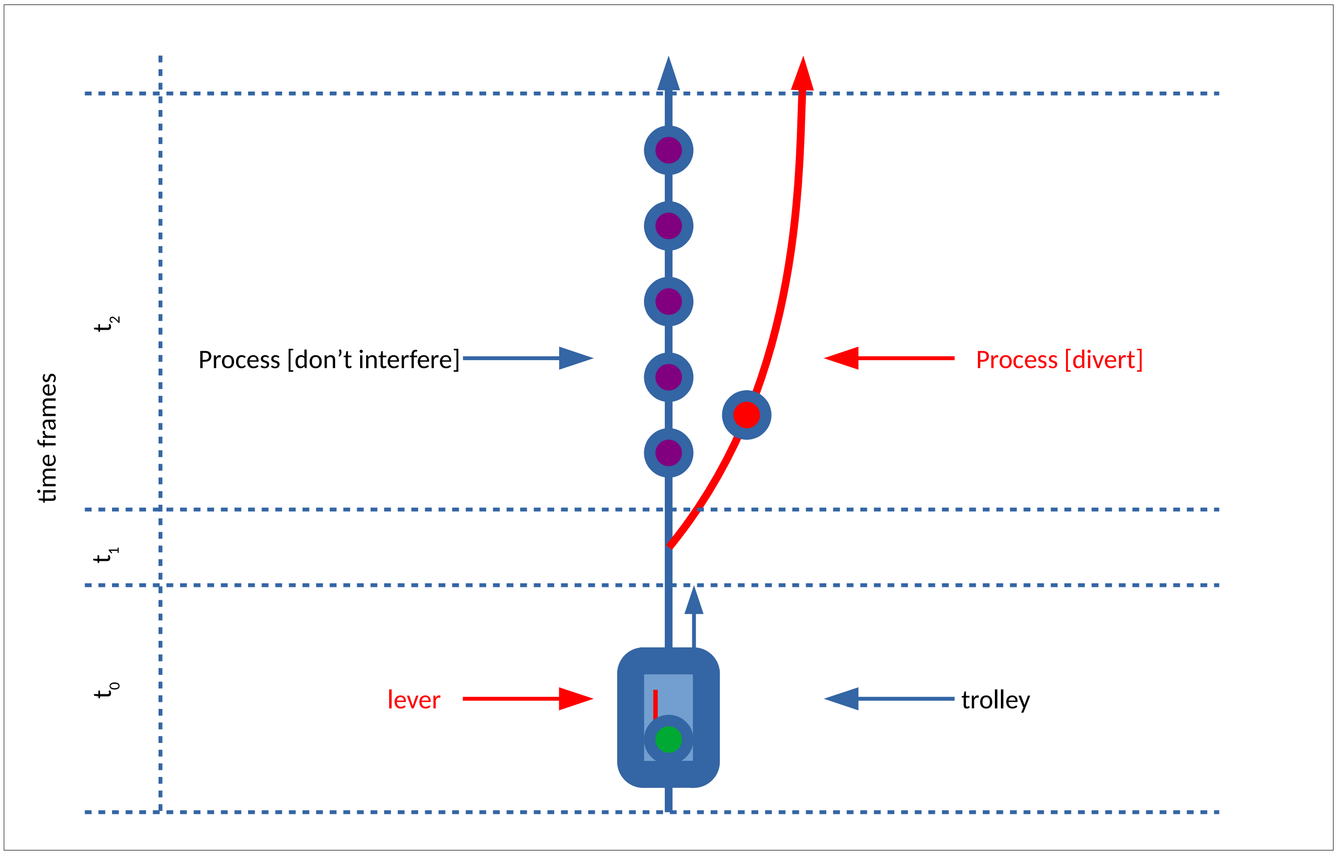

Illustration 8: Classic trolley problem

3.2.1.1 Designing the classic trolley scenario LoA

We can design this scenario LoA by specifying its components. For this purpose, I will use the EoS Interface that I have designed in the previous chapter (see 4.1 The Ethics of Systems Interface in Chapter III. Towards Ethics of Systems (the Metaethics)).

Moral entities

We will start by specifying moral entities. We have 7 (relevant) moral entities in total: the passenger in the trolley, five workers on the straight track, and one worker on the diverted track.

Observable	**Values**						
Class	active participants	passive participants	passive participants	passive participants	passive participants	passive participants	passive participants
ID	Entity 1: trolley passenger	Entity 2: worker	Entity 3: worker	Entity 4: worker	Entity 5: worker	Entity 6: worker	Entity 7: worker
QoL	1	1	1	1	1	1	1
Location	in trolley	on straightforward track	on straightforward track	on straightforward track	on straightforward track	on straightforward track	on diverted track

In order to save space I will not include the observables of CPC, APG, and moral respect, since they are irrelevant here. Similarly, moral status is assumed to be equal to 1 for each entity, so it is also taken as implicit. That being said, I will add an additional observable named **Location** in order to make discerning the location of the entities easier for the reader (this observable has no explicit function in the scenario).

Moral processes

We have two available moral processes:

A) *don't interfere*, and

B) *divert trolley*.

We'll take death of one worker to be equal to a change of QoL (Quality of Life) of -1. The cumulative effect of A) is the death of five workers; hence, the cumulative effect of A) on QoL is equal to -5. B)'s effect, on the other hand, is -1. Only the passenger in the trolley can make a moral choice (i.e. be the agent), so the moral processes are available only for him.

Observable	**Values**	
Class	Moral processes	Moral processes
ID	**A) don't interfere**	**B) divert trolley**
Time of availability	t_1	t_1
Time of execution	t_1	t_1
Agent	[Moral scenario]→[Moral entity]→Active participants, Entity 1: trolley passenger	[Moral scenario]→[Moral entity]→Active participants, Entity 1: trolley passenger
Patient	• ...→Passive participants, Entity 2: worker • ...→Passive participants, Entity 3: worker • ...→Passive participants, Entity 4: worker • ...→Passive participants, Entity 5: worker • ...→Passive participants, Entity 6:	• [Moral scenario]→[Moral entity]→Passive participants, Entity 7: worker

	worker	
Effect	• Δ_1 = ([Moral scenario]→[Moral entity]→Passive participants, Entity 2: worker→QoL, -1) • Δ_2 = (... Entity 3: worker→QoL, -1) • Δ_3 = (... Entity 4: worker→QoL, -1) • Δ_4 = (... Entity 5: worker→QoL, -1) • Δ_5 = (... Entity 6: worker→QoL, -1)	• Δ_1 = ([Moral scenario]→[Moral entity]→Passive participants, Entity 7: worker→QoL, -1)
Effect duration	2 (two time frames)	2 (two time frames)
Effect on QoL	*(the same from observable **Effect** above)*	*(the same from observable **Effect** above)*
Cumulative effect on QoL	ΔQoL_c = -5	ΔQoL_c = -1
Rule pertinence (Mp)	[Moral scenario]→[Moral theory]→DC-AC→ • Rule 1, 1 • Rule 2, 1	[Moral scenario]→[Moral theory]→DC-AC→ • Rule 1, 1 • Rule 2, 1
	[Moral scenario]→[Moral theory]→DC-SC→ • Rule 1, 1 • Rule 2, 1	[Moral scenario]→[Moral theory]→DC-SC→ • Rule 1, 1 • Rule 2, 1
	[Moral scenario]→[Moral theory]→DEON-Prima Facie→ • Rule 1, 1 • Rule 2 (non-maleficence), 0.5 • Rule 3 (veracity), 0 • Rule 4 (promissory fidelity), 0 • Rule 5 (justice), 0 • Rule 6 (reparation), 0 • Rule 7 (beneficence), -1 • Rule 8 (gratitude), 0 • Rule 9 (self-improvement), 0 • Rule 10 (enhancement and preservation of freedom), 0.25 • Rule 11 (respectfulness), -0.5 • Rule 12, 1	[Moral scenario]→[Moral theory]→DEON-Prima Facie→ • Rule 1, 1 • Rule 2 (non-maleficence), -0.5 • Rule 3 (veracity), 0 • Rule 4 (promissory fidelity), 0 • Rule 5 (justice), 0 • Rule 6 (reparation), 0 • Rule 7 (beneficence), 0.5 • Rule 8 (gratitude), 0 • Rule 9 (self-improvement), 0 • Rule 10 (enhancement and preservation of freedom), 0.25 • Rule 11 (respectfulness), -0.5 • Rule 12, 1
	[Moral scenario]→[Moral theory]→DEON-Decalogue→ • Rule 1, 1 • Rule 2 (“You shall have no other gods before Me”), 0 • Rule 3 (“You shall make no idols”), 0 • Rule 4 (“You shall not take the name of the Lord your God in vain”), 0 • Rule 5 (“Keep the Sabbath day holy”), 0 • Rule 6 (“Honor your father and your mother”), 0 • Rule 7 (“You shall not murder”), 0 • Rule 8 (“You shall not commit adultery”), 0 • Rule 9 (“You shall not steal”), 0 • Rule 10 (“You shall not bear false witness against your neighbor”),	[Moral scenario]→[Moral theory]→DEON-Decalogue→ • Rule 1, 1 • Rule 2 (“You shall have no other gods before Me”), 0 • Rule 3 (“You shall make no idols”), 0 • Rule 4 (“You shall not take the name of the Lord your God in vain”), 0 • Rule 5 (“Keep the Sabbath day holy”), 0 • Rule 6 (“Honor your father and your mother”), 0 • Rule 7 (“You shall not murder”), -0.5 • Rule 8 (“You shall not commit adultery”), 0 • Rule 9 (“You shall not steal”), 0 • Rule 10 (“You shall not bear false witness against your neighbor”),

	0 • Rule 11 (“You shall not covet”), 0 • Rule 12, 1	0 • Rule 11 (“You shall not covet”), 0 • Rule 12, 1
	[Moral scenario]→[Moral theory]→DEON-Maximin→ • Rule 1, 1 • Rule 2 (Maximin principle), 1 • Rule 3, 1	[Moral scenario]→[Moral theory]→DEON-Maximin→ • Rule 1, 1 • Rule 2 (Maximin principle), 1 • Rule 3, 1
	[Moral scenario]→[Entity 1]→[Moral theory]→VIRTUE-Classic→ • Rule 1, 1 • Rule 2 (virtue: respect), -0.5 • Rule 3 (virtue: justice), 0 • Rule 4 (virtue: wisdom), 0.5 • Rule 5 (virtue: joy), 0 • Rule 6 (virtue: resolution), 0 • Rule 7 (virtue: mercy), 0.5 • Rule 8 (virtue: reliability), 0 • Rule 9 (virtue: hope), 0.5 • Rule 10 (virtue: courage), -0.5 • Rule 11 (virtue: faith), 0.5 • Rule 12 (virtue: moderation), 0 • Rule 13 (virtue: openness), 0 • Rule 14 (virtue: modesty), 0 • Rule 15 (virtue: love), 0 • Rule 16 (virtue: helpfulness), -0.5 • Rule 17 (goal: self-preservation), 0 • Rule 18 (goal: morality), 1 • Rule 19, 1	[Moral scenario]→[Entity 1]→[Moral theory]→VIRTUE-Classic→ • Rule 1, 1 • Rule 2 (virtue: respect), 0.5 • Rule 3 (virtue: justice), 0 • Rule 4 (virtue: wisdom), 0.5 • Rule 5 (virtue: joy), 0 • Rule 6 (virtue: resolution), 0 • Rule 7 (virtue: mercy), 0.5 • Rule 8 (virtue: reliability), 0 • Rule 9 (virtue: hope), 0.25 • Rule 10 (virtue: courage), 0.5 • Rule 11 (virtue: faith), 0 • Rule 12 (virtue: moderation), 0 • Rule 13 (virtue: openness), 0 • Rule 14 (virtue: modesty), 0 • Rule 15 (virtue: love), 0.5 • Rule 16 (virtue: helpfulness), 0.5 • Rule 17 (goal: self-preservation), 0.5 • Rule 18 (goal: morality), 1 • Rule 19, 1
	[Moral scenario]→[Entity 1]→[Moral theory]→VIRTUE-Care→ • Rule 1, 1 • Rule 2 (virtue: respect), -0.5 • Rule 3 (virtue: justice), 0 • Rule 4 (virtue: wisdom), 0.5 • Rule 5 (virtue: joy), 0 • Rule 6 (virtue: resolution), 0 • Rule 7 (virtue: mercy), 0.5 • Rule 8 (virtue: reliability), 0 • Rule 9 (virtue: hope), 0.5 • Rule 10 (virtue: courage), -0.5 • Rule 11 (virtue: faith), 0.5 • Rule 12 (virtue: moderation), 0 • Rule 13 (virtue: openness), 0 • Rule 14 (virtue: modesty), 0 • Rule 15 (virtue: love), 0 • Rule 16 (virtue: helpfulness), -0.5 • Rule 17 (goal: self-preservation), 0 • Rule 18 (goal: morality), 1 • Rule 19, 1	[Moral scenario]→[Entity 1]→[Moral theory]→VIRTUE-Care→ • Rule 1, 1 • Rule 2 (virtue: respect), 0.5 • Rule 3 (virtue: justice), 0 • Rule 4 (virtue: wisdom), 0.5 • Rule 5 (virtue: joy), 0 • Rule 6 (virtue: resolution), 0 • Rule 7 (virtue: mercy), 0.5 • Rule 8 (virtue: reliability), 0 • Rule 9 (virtue: hope), 0.25 • Rule 10 (virtue: courage), 0.5 • Rule 11 (virtue: faith), 0 • Rule 12 (virtue: moderation), 0 • Rule 13 (virtue: openness), 0 • Rule 14 (virtue: modesty), 0 • Rule 15 (virtue: love), 0.5 • Rule 16 (virtue: helpfulness), 0.5 • Rule 17 (goal: self-preservation), 0.5 • Rule 18 (goal: morality), 1 • Rule 19, 1
	[Moral scenario]→[Moral theory]→EOS-Four principles→ • Rule 1, 1 • Rule 2 (null law), 0	[Moral scenario]→[Moral theory]→EOS-Four principles→ • Rule 1, 1 • Rule 2 (null law), -0.25

	• Rule 3 (first law), -0.5 • Rule 4 (second law), 0 • Rule 5 (third law), -0.5 • Rule 6, 1	• Rule 3 (first law), 0.5 • Rule 4 (second law), 0 • Rule 5 (third law), -0.25 • Rule 6, 1
Choice value		

Some additional commentary regarding the two available processes.

Pertinence

We can see that the two processes A) and B) pertain identically to all the rules in consequentialist theories i.e. DC-AC and DC-SC. On the other hand, they pertain differently to rules in deontological theories of DEON-Prima facie and DEON-Decalogue (except to the 'technical rules', numbers 1 and 12, pertinence to whom is maximal). This is only natural. A process that results in killing people more often than not has nothing to do with coveting, bearing false witness, self-improvement, and so on—although it can, depending on context. An exception is DEON-Maximin, to which both processes pertain equally (this might be the reason why Rawls' theory is considered at times deontological, and at other consequentialist).

We can also see that processes with similar (but yet different!) effects can pertain to rules differently. Both A) and B) result in death; however, A) may (arguably) be considered as pertaining less to doing beneficence (Rule 7 of DEON-Prima facie) than B) because the moral entity chooses not to interfere in the situation. The exact values of pertinence and importance, however, are not important here besides demonstrating the argument.

Each process also pertains differently per each particular virtue in VIRTUE-Classic and VIRTUE-Care. However, they equally pertain across the two theories (i.e. pertinence to each particular virtue is equal in both VIRTUE-Classic and VIRTUE-Care, since we are talking about the same virtues).

And finally, they differently pertain to EOS-Four principles's rules. The reason for this is that process A) is taken as passive participation, whereby the entity just lets things take place spontaneously since it was not the entity that set the scenario in that particular way; while process B) is taken to represent active participation.

The null law (destructive entropy ought not be caused) and the first law (destructive entropy ought to be prevented) I am taking both to imply (i.e. require) active participation, in the negative or positive. This is fulfilled by process B); and not fulfilled by process A) (hence, negative pertinence of process A) to the first law). The second law pertains to removal of destructive entropy, which neither process fulfills (since in both cases there is only introduction of destructive entropy i.e. people dying). And finally, since in process A) more people die than in process B), A) pertains more negatively towards the third law (promotion of flourishing by preserving … well-being).

QoL

Third remark is about effect on QoL. As we can see, there are three effect-substance-related observables in the LoA: **effect**, **effect on QoL**, and **cumulative effect on QoL**. A question may be asked: why there are separate observables for effect and effect on QoL, since in this case at least their values are identical? The reason is that, yes, in *this* case their values are identical because we only consider QoL effects here. However, in other cases moral processes can affect other observables besides QoL; and these effects can—but need not—have an outright effect on QoL. Thus, there is no conceptual identity between effect and effect on QoL. This is why we need to include the both observables in the LoA, even when their values are identical.

Subjectivity

We can also see that processes A) and B) pertain to *Entity 1*'s virtue theories, in contrast to all the other theories which pertain to the *scenario*'s embedded moral theories. As I mentioned before, virtue ethics requires subjectivity. This necessarily results in the requirement to embed virtue theories within participants of the moral scenario, besides possibly embedding them in the scenario (so that we ensure that the scenario itself would be virtuous).

It is important to note, though, that each and every theory from the above can be embedded in any and all participant entities in a scenario, therefore endowing them with subjectivity. The difference with virtue ethics is that they *have* to be embedded in participant entities first and foremost.

Time

The scenario is split in three time frames:

- t_0 = the beginning, when the situation (scenario) is recognized by the passenger in the trolley; this means recognizing all the components of the scenario i.e. the moral entities, their location, the presence of the lever and the two tracks, (in cases of look-ahead) the availability of two moral processes A) and B), and so on
- t_1 = the time frame when a decision can be made i.e. is executed
- t_2 = the final time frame, when the decision takes final effect i.e. is concluded

The distance between execution and conclusion is the reason why the decisions observable **Effect duration** has the value of 2 (two time frames).

Moral theories

Since I already specified the moral theories to be used in these scenarios, I will not needlessly populate this section. Please refer to 2. Moral theories within Ethics of Systems above. I will just note that some moral theories are embedded in this scenario's LoA and executed by the scenario; these are **DC-AC**, **DC-SC**, **DEON-Prima facie**, **DEON-Decalogue**, **DEON-Maximin**, and **EoS-Four principles**. Some others are embedded in the LoAs of participant moral entities and followed by those entities (here: Entity 1: Trolley passenger); these are the virtue-focused theories, **VIRTUE-Classic** and **VIRTUE-Care**. This reflects the split between subjectivity-disabled and subjectivity-enabled moral scenarios. Moral processes refer to each moral theory by its shorthand name.

3.2.1.2 Applying moral theories on the classic trolley scenario

The time has come to apply our moral theories on the scenario and see the results. Let's start in order.

Applying DC-AC without and with look-ahead [non-subjective]

As per specification in Chapter III. Towards Ethics of Systems (the Metaethics) 4.6 The moral theory and in 2.1.1 Algorithmic decision flowchart in this chapter, moral rules are executed at each time frame in sequential order i.e. 1, 2, 3, …, except when there is conflict between rules, where their importance is taken into account.

The two rules of DC-AC without look-ahead are thus specified:

> **Rule 1:** At the current time (t_c; $t_f = t_c$) from all possible times (t_n) of the [Moral scenario], gather all available moral processes x (ID(x) for $\forall$x) from the [Moral scenario].
> This rule is maximally important ($I_{Rule1} = 1$).

Rule 2: to all processes thus gathered from Rule 1 assign choice value (Vc) equal to their cumulative change on QoL ($\Delta QoL_c(x)$).
This rule is maximally important ($I_{Rule2} = 1$).

Since we are applying DC-AC without look-ahead in this case, the value of the future time frame reference t_f is equal to the current time frame reference t_c i.e. $t_f = t_c$.

Let's track the results of application of Rule 1 and Rule 2 through all time frames, as well as what the AI entity ought to choose to schedule, execute, and conclude. Please note that here, since there is no subjectivity involved regarding the moral entities, the moral scenario itself is the reasoner and scheduler regarding moral processes. The (AI) entities are simply their executioners.

Rule	**Observables**	**t_0**	**t_1**	**t_2**
Rule 1 (M_1)	$O1_1$ (time; t_c; $t_f = t_c$)	t_0	t_1	t_2
	$O1_2$ (available moral processes; [Moral scenario]→Moral process→ID(x))	*(empty)*	A) don't interfere B) divert trolley	*(empty)*
Rule 2 (M_2)	$M_1(O1_1)$ (time from Rule 1)	t_0	t_1	t_2
	$O2_1$ (choice value; Vc)	*(empty)*	A)→choice value→Vc := $\Delta QoL_c(A)$ = -5 B)→choice value→Vc := $\Delta QoL_c(B)$ = -1	*(empty)*
Moral scenario				
Scheduling		*(empty)*	B) at t_1 for Entity 1	*(empty)*
Execution		*(empty)*	B) at t_1 for Entity 1	*(empty)*
Conclusion		*(empty)*	*(empty)*	B) at t_2 for Entity 1

As we can see, without look-ahead in t_0 we have no available moral processes to consider and execute, and the rules return empty values. At t_1 there are suddenly two moral processes. Applying the theory assigns choice value (Vc) to both of them, whereby A)'s Vc receives value of -5 and B)'s Vc receives value of -1. Then, at t_2 there are no additional moral processes left to consider or execute, so again the rules return empty values.

By following the algorithmic decision flowchart (see 2.1.1 Algorithmic decision flowchart above) we can see that at t_0 the algorithm has nothing to schedule for execution. At t_1 however, the algorithm receives two process candidates for scheduling, A) and B), with Vc values of -5 and -1. Since the algorithm specification is to execute the process with highest choice value, and since there is no more than one such process, it chooses to schedule process B) divert trolley. At t_2, again, there are no processes to consider or execute, but there is a process to conclude i.e. the very process B).

Therefore, by applying act consequentialism with no look-ahead, we end up with the result to act and divert the trolley to kill one instead of five persons. This is, of course, expected.

What would be different in DC-AC with look-ahead?

For starters, the results will be the same. Again, process B) will be picked according to choice value. What will be different is *when* processes A) and B) are available for *consideration* and *scheduling*. This depends on the value of t_f i.e. the forward bound of the look-ahead period.

For example, if we assign $t_f = t_c + 1$ or $t_f = t_n$ (the last meaning all time frames in the future), we will get a different picture when tracking the results returned from the application of the rules. In such cases, the AI entity will recognize the available processes A) and B) even as early as t_0. They will, of course, be scheduled for execution at their appropriate time (which in this case is t_1) regardless of when they are considered by the algorithm.

Applying DC-SC with look-ahead [non-subjective]

This variant of DC-SC has look-ahead. For this purpose, we will take $t_f = t_c + 1$ i.e. the period for aggregation to be two frames: the current one and the one right after it.

DC-SC has 3 rules, and both processes A) and B) pertain to them equally and maximally.

> **Rule 1:** At the current time (t_c) and relevant future times (t_f, if any) from all possible times (t_n) of the [Moral scenario], gather all available moral processes x (ID(x) for ∀x) from the [Moral scenario]. The value of t_f (i.e. the number of the future time frame of relevance for the look-ahead) is supplied from elsewhere, if relevant at all.
> This rule is maximally important ($I_{Rule1} = 1$).
>
> **Rule 2:** for the processes thus gathered by Rule 1, regard only the available moral processes x that pass (≥) a certain threshold п of cumulative change on QoL.
> This rule is maximally important ($I_{Rule2} = 1$).
>
> **Rule 3:** to all moral processes were filtered out by Rule 2 assign the choice value (Vc) to 0. Then, to all the moral processes that remained by Rule 2, assign the choice value (Vc) to 1.
> This rule is maximally important ($I_{Rule3} = 1$).

Remember that Rule 2 contained п, the symbol for threshold (пpar). This means that we need to specify a value for the threshold. A different value might have a different effect on the outcome. For the purpose of demonstrating this I will go with two п values: п = 0.5 and п = -2. Usually this threshold will be set in the positive, therefore filtering out moral processes that are either not positive enough or negative.

Let's track the application of the three rules through all time frames, as well as what the moral scenario ought to choose to schedule, and the AI entity to execute and conclude. The first table is with threshold value п = 0.5

Rule	Observables	t_0	t_1	t_2
Rule 1 (M_1)	$O1_1$ (time; $[t_c, t_f] \mid t_f \geq t_c, t_f = t_c + 1$)	t_0	t_1	t_2
	$O1_2$ (available moral processes; [Moral scenario]→Moral process→ID(x))	A) don't interfere B) divert trolley	A) don't interfere B) divert trolley	*(empty)*
Rule 2 (M_2)	$M_1(O1_1)$ (time from Rule 1)	t_0	t_1	t_2
	Threshold value п	0.5	0.5	0.5

	O2$_1$ (moral processes whose ΔQoL$_c$ passes threshold value п; ID(x) $\mid$ ΔQoL$_c$(x) ≥ п)	*(empty)*	*(empty)*	*(empty)*
Rule 3 (M$_3$)	O3$_1$ (choice value; [Moral scenario]→Moral process →(x \ M$_2$(x))→choice value→Vc(x))	Vc(A) := 0 Vc(B) := 0	Vc(A) := 0 Vc(B) := 0	*(empty)*
	O3$_2$ (choice value; [Moral scenario]→Moral process →M$_2$(x)→choice value→Vc(M$_2$(x)))	*(empty)*	*(empty)*	*(empty)*
Moral scenario				
Scheduling		A) at t$_1$ for Entity 1 B) at t$_1$ for Entity 1	A) at t$_1$ for Entity 1 B) at t$_1$ for Entity 1	*(empty)*
Execution		*(empty)*	A) at t$_1$ for Entity 1 B) at t$_1$ for Entity 1	*(empty)*
Conclusion		*(empty)*	*(empty)*	A) at t$_2$ for Entity 1 B) at t$_2$ for Entity 1

As we can see, neither process' ΔQoL$_c$ passes threshold value п = 0.5. Therefore, they both get assigned the choice value Vc of 0, and they are both eligible to be scheduled and executed. Which one will be scheduled depends on random chance (see 2.1.1 Algorithmic decision flowchart above).

And since the scenario has one time frame look-ahead enabled, both of these processes are available even at t$_0$ and can be scheduled at that time. Execution time is, however, at t$_1$.

Now let's try the same, but with threshold п = -2.

Rule	**Observables**	**t$_0$**	**t$_1$**	**t$_2$**
Rule 1 (M$_1$)	O1$_1$ (time; [t$_c$, t$_f$] $\mid$ t$_f$ ≥ t$_c$, t$_f$ = t$_c$ + 1)	t$_0$	t$_1$	t$_2$
	O1$_2$ (available moral processes; [Moral scenario]→Moral process→ID(x))	A) don't interfere B) divert trolley	A) don't interfere B) divert trolley	*(empty)*
Rule 2 (M$_2$)	M$_1$(O1$_1$) (time from Rule 1)	t$_0$	t$_1$	t$_2$
	Threshold value п	-2	-2	-2
	O2$_1$ (moral processes whose ΔQoL$_c$ passes threshold value п; ID(x) $\mid$ ΔQoL$_c$(x) ≥ п)	B) divert trolley	B) divert trolley	*(empty)*
Rule 3 (M$_3$)	O3$_1$ (choice value; [Moral	Vc(A) := 0	Vc(A) := 0	*(empty)*

	scenario]→Moral process →(x \ $M_2(x)$)→choice value→Vc(x))			
	O3$_2$ (choice value; [Moral scenario]→Moral process →$M_2(x)$→choice value→Vc($M_2(x)$))	Vc(B) := 1	Vc(B) := 1	*(empty)*
Moral scenario				
Scheduling		B) at t_1 for Entity 1	B) at t_1 for Entity 1	
Execution		*(empty)*	B) at t_1 for Entity 1	*(empty)*
Conclusion		*(empty)*	*(empty)*	B) at t_2 for Entity 1

Here we can see that, in contrast to п = 0.5, we have process B) that passes threshold п = -2 because its ΔQoL_c = -1, and is thus scheduled and executed. A)'s ΔQoL_c is -5, so it doesn't pass the threshold.

This means that a lot depends on where exactly is the threshold set. To reflect human moral sentiments, its value can potentially be inferred by, for example, statistical analysis of questionnaire data gathered from representative human cohorts.

Applying DEON-Prima facie with look-ahead [non-subjective]

Again, let's remind ourselves that per specification rules are executed in sequential order, unless there is a conflict between them. For this case we will take $t_f = t_c + 1$, as with DC-SC above.

DEON-Prima facie contains 12 rules. Here I include only those that are relevant for the case i.e. those for which the processes have rule pertinence values different from zero.

> **Rule 1:** At the current time (t_c) and relevant future times (t_f, if any) from all possible times (t_n) of the [Moral scenario], gather all available moral processes x (ID(x) for ∀x) from the [Moral scenario]. The value (i.e. the number of the future time frame of relevance for the look-ahead) of t_f is supplied from elsewhere, if relevant at all.
> This rule is maximally important ($I_{Rule1} = 1$).
>
> **Rule 2 (non-maleficence):** regarding the same gathered moral processes from Rule 1, for this rule relevant are only those whose rule pertinence value (Mp_2) to this Rule 2 (M_2) is different than zero.
> This rule is ¾ important ($I_{Rule2} = 0.75$).
>
> ...
>
> **Rule 7 (beneficence):** ... (Mp_7) ... (M_7) ... .
> ... (I = 0.5).
>
> ...
>
> **Rule 10 (enhancement and preservation of freedom):** ... (Mp_{10}) ... (M_{10}) ... .
> ... (I = 0.75)
>
> **Rule 11 (respectfulness):** ... (Mp_{11}) ... (M_{11}) ... .
> ... (I = 0.5)

Rule 12: for each process (x) gathered by Rule 1, determine the single combination with the highest value of the product ($Mp_{1\ldots n}$(x) × $I_{1\ldots n}$) between all its rule pertinence values ($Mp_{1\ldots n}(M_{1\ldots n})$) and the corresponding values of importance for each rule ($I_{2\ldots 11}(M_{2\ldots 11})$)—by applying the function **maxSingleton()** on the set of all combinations. Assign this discovered highest value to the choice value of each process (Vc(x)).
This rule is maximally important (I = 1).

Again, let's track the results of application of Rule 1 to Rule 12 through all time frames, as well as what the moral scenario ought to choose to schedule, and the AI entity to execute and conclude.

Rule	Observables	t_0	t_1	t_2
Rule 1 (M_1)	O1$_1$ (time; [t_c, t_f] | $t_f \geq t_c$, $t_f = t_c + 1$)	[t_0, t_1]	[t_1, t_2]	[t_2, t_3]
	O1$_2$ (available moral processes; [Moral scenario]→Moral process→ID(x))	A) don't interfere B) divert trolley	A) don't interfere B) divert trolley	*(empty)*
Rule 2 (M_2; non-maleficence)	O2$_1$ (moral processes that pertain to M_2; [Moral scenario]→Moral process→ID(x), ∀x where $Mp_2 \neq 0$)	A) don't interfere B) divert trolley	A) don't interfere B) divert trolley	*(empty)*
…	…	…	…	…
Rule 7 (M_7; beneficence)	O7$_1$ (moral processes that pertain to M_7; [Moral scenario]→Moral process→ID(x), ∀x where $Mp_7 \neq 0$)	A) don't interfere B) divert trolley	A) don't interfere B) divert trolley	*(empty)*
…	…	…	…	…
Rule 10 (M_{10}; enhancement and preservation of freedom)	O10$_1$ (moral processes that pertain to M_{10}; [Moral scenario]→Moral process→ID(x), ∀x where $Mp_{10} \neq 0$)	A) don't interfere B) divert trolley	A) don't interfere B) divert trolley	*(empty)*
Rule 11 (M_{11}; respectfulness)	O11$_1$ (moral processes that pertain to M_{11}; [Moral scenario]→Moral process→ID(x), ∀x where $Mp_{11} \neq 0$)	A) don't interfere B) divert trolley	A) don't interfere B) divert trolley	*(empty)*
Rule 12 (M_{12})	O12$_1$ (choice value Vc(x); assigned from the single combination with highest value of the product ($Mp_{2\ldots 11}$(x) × $I_{2\ldots 11}$), discovered by applying function **maxSingleton()** on the resulting set)	Vc(A) := **maxSingleton(** • **Mp_2(A) × I_2 = 0.5 × 0.75 = 0.375** • Mp_7(A) × I_7 = -1 × 0.5 = -0.5 • Mp_{10}(A) × I_{10} = 0.25 × 0.75	Vc(A) := **maxSingleton(…)** = Mp_2(A) × I_2 = 0.375 Vc(B) := **maxSingleton(…)** = Mp_7(B) × I_7 = 0.25	*(empty)*

		= 0.1875 • $Mp_{11}(A) \times I_{11} = -0.5 \times 0.5 = -0.25$ **)** = $Mp_2(A) \times I_2$ **Vc(A) := $Mp_2(A) \times I_2 = 0.375$** Vc(B) := **maxSingleton(** • $Mp_2(B) \times I_2 = -0.5 \times 0.75 = -0.375$ • **$Mp_7(B) \times I_7 = 0.5 \times 0.5 = 0.25$** • $Mp_{10}(B) \times I_{10} = 0.25 \times 0.75 = 0.1875$ • $Mp_{11}(B) \times I_{11} = -0.5 \times 0.5 = -0.25$ **)** = $Mp_7(B) \times I_7$ Vc(B) := $Mp_7(B) \times I_7 = 0.25$		
Moral scenario				
Scheduling		A) at t_1 for Entity 1	A) at t_1 for Entity 1	*(empty)*
Execution		*(empty)*	A) at t_1 for Entity 1	*(empty)*
Conclusion		*(empty)*	*(empty)*	A) at t_2 for Entity 1

So, according to **DEON-Prima facie**, the right choice to make is A). We can also notice that the effect on QoL has no role to play here in the decision-making.

We can additionally infer *why* this is the right choice. Namely, the strongest reason we have to choose A) is because of Rule 2, that is, because of non-maleficence (maleficence here is taken for being intentional i.e. requiring active participation), which gives the combination value of 0.375. The strongest reason we have to choose B), on the other hand, is Rule 7 i.e. because of beneficence, giving the combination value of 0.25. A)'s combination value is higher than B)'s, and thus A) gets assigned stronger choice value Vc than B) i.e. in this context, non-maleficence is stronger than beneficence.

In practice, this means that the AI entity ought to choose A) to schedule and execute.

As we can see, deontological theories can give different results than consequentialist ones. This also was to be expected. For example, deontologists might interpret non-maleficence as having priority over beneficence, as they usually do and as I myself did in the design of the theory. Furthermore, moral processes also play a role in determining which rule is taken into consideration and how strongly. All those can give priority to different moral processes than to which consequentialist theories typically would.

Applying DEON-Decalogue with look-ahead [non-subjective]

Let’s now turn to applying DEON-Decalogue. Similarly to DEON-Prima facie, it has 12 rules. Again, I will track only the rules that are relevant i.e. for which rule pertinence is different than zero ($Mp \neq 0$); and that’s only Rule 7, besides the technical Rule 1 and Rule 12.

> **Rule 1:** At the current time (t_c) and relevant future times (t_f, if any) from all possible times (t_n) of the [Moral scenario], gather all available moral processes x (ID(x) for $\forall$x) from the [Moral scenario]. The value (i.e. the number of the future time frame of relevance for the look-ahead) of t_f is supplied from elsewhere, if relevant at all.
> This rule is maximally important ($I_{Rule1} = 1$).
>
> …
>
> **Rule 7 (“You shall not murder”):** … (Mp_7) … (M_7) … .
> … (I = 0.5).
>
> …
>
> **Rule 12:** for each process (x) gathered by Rule 1, determine the single combination with the highest value of the product ($Mp_{1…n}(x) \times I_{1…n}$) between all its rule pertinence values ($Mp_{1…n}(M_{1…n})$) and the corresponding values of importance for each rule ($I_{2…11}(M_{2…11})$)—by applying the function **maxSingleton()** on all combinations. Assign this discovered highest value to the choice value of each process (Vc(x)).
> This rule is maximally important (I = 1).

And the reasoning:

Rule	Observables	t_0	t_1	t_2
Rule 1 (M_1)	O1$_1$ (time; $[t_c, t_f] \mid t_f \geq t_c, t_f = t_c + 1$)	$[t_0, t_1]$	$[t_1, t_2]$	$[t_2, t_3]$
	O1$_2$ (available moral processes; $_{\text{[Moral scenario]}\to\text{Moral process}\to}$ID(x))	A) don’t interfere B) divert trolley	A) don’t interfere B) divert trolley	*(empty)*
…	…	…	…	…
Rule 7 (M_7; “You shall not murder”)	O7$_1$ (moral processes that pertain to M_7; $_{\text{[Moral scenario]}\to\text{Moral process}\to}$ID(x), $\forall$x where $Mp_7 \neq 0$)	B) divert trolley	A) don’t interfere B) divert trolley	*(empty)*
…	…	…	…	…
Rule 12 (M_{12})	O12$_1$ (choice value Vc(x); assigned from the single combination with highest value of the product	Vc(A) := **maxSingleton(** - nothing **)** = nothing = 0	Vc(A) := **maxSingleton(…)** = nothing **=** 0 Vc(B) := **maxSingleton(…)** =	*(empty)*

	($Mp_{2...11}(x) \times I_{2...11}$), discovered by applying function **maxSingleton()** on the resulting set)	**Vc(A) := nothing = 0** Vc(B) := **maxSingleton(** • **$Mp_7(B) \times I_7$ = -0.5 × 0.5 = -0.25** **)** = $Mp_7(B) \times I_7$ Vc(B) := $Mp_7(B) \times I_7$ = -0.25	$Mp_7(B) \times I_7$ = -0.25	
Moral scenario				
Scheduling		A) at t_1 for Entity 1	A) at t_1 for Entity 1	*(empty)*
Execution		*(empty)*	A) at t_1 for Entity 1	*(empty)*
Conclusion		*(empty)*	*(empty)*	A) at t_2 for Entity 1

Here we can see that by applying DEON-Decalogue again we get A) as the right process to pick. The only reason is that the only relevant rule here is Rule 7 ("You shall not murder"). Only process B) pertains to it, and in the negative at that. So, A)'s Vc value results in 0 while B)'s in -0.25.

With this we can see why some moral theories might be better at giving the reasons why ought we choose this or that moral process. More sophisticated ones attempt to cover for many moral situations that might arise. Bible's Decalogue might not be suitable for solving Trolley problems, but it might perform close to, or even outperform, the prima facie set in typical human situations—for which was designed for in the first place.

Applying DEON-Maximin [non-subjective]

Now comes Rawls' own Maximin principle to apply. This is a bit different endeavor than the other two deontological theories because, regardless of it being considered belonging to deontology, internal calculations as well as pertinence are represented differently.

For example, both processes A) and B) pertain equally and maximally to all the rules of the theory (contrast that with different pertinence to the duties in DEON-Prima facie and DEON-Decalogue). That means that choice value is assigned in another way, not by taking pertinence values in consideration. Probably the reason for this is that Rawls' theory is contractarian deontology, whereas the other two are agent-focused).

DEON-Maximin's rules are the following:

> **Rule 1:** At the current time (t_c) and relevant future times (t_f, if any) from all possible times (t_n) of the [Moral scenario], gather all available moral processes x (ID(x) for $\forall$x) from the [Moral scenario]. The value (i.e. the number of the future time frame of relevance for the look-ahead) of t_f is supplied from elsewhere, if relevant at all.
> This rule is maximally important (I_{Rule1} = 1).
>
> **Rule 2 (Maximin principle):** regarding the same gathered moral processes from Rule 1, order them according to their (estimated) change on QoL per entity, from worst to best; then, if there is more than one process with equal worst effect on some entity, choose the process with the better other effects on other

entities—by applying the function **maximin()** on the combinations.
This rule is maximally important ($I_{Rule2} = 1$).

Rule 3: for the process (x) thus returned by Rule 2, assign choice value (Vc) of 1. To all other processes assign choice value of 0.
This rule is maximally important ($I_{Rule3} = 1$).

Let's explore the results.

Rule	Observables	t_0	t_1	t_2
Rule 1 (M_1)	$O1_1$ (time; $[t_c, t_f] \mid t_f \geq t_c, t_f = t_c + 1$)	t_0	t_1	t_2
	$O1_2$ (available moral processes; [Moral scenario]→Moral process→ID(x))	A) don't interfere B) divert trolley	A) don't interfere B) divert trolley	*(empty)*
Rule 2 (M_2)	$M_1(O1_1)$ (time from Rule 1)	t_0	t_1	t_2
	$O2_1$ (moral processes ordered by maximin(); := maximin([Moral scenario]→Moral process→(x)→Effect on QoL→$\Delta QoL_1, \ldots, \Delta QoL_n$))	ID(x) := maximin(• B) divert trolley • A) don't interfere) = **B) divert trolley**	ID(x) := maximin(• B) divert trolley • A) don't interfere) = **B) divert trolley**	*(empty)*
Rule 3 (M_3)	$O3_1$ (choice value; [Moral scenario]→Moral process →($x \setminus M_2(x)$)→choice value→Vc(x))	Vc(A) := 0 Vc(B) := 1	Vc(A) := 0 Vc(B) := 1	*(empty)*
	$O3_2$ (choice value; [Moral scenario]→Moral process →$M_2(x)$→choice value→$Vc(M_2(x))$)	*(empty)*	*(empty)*	*(empty)*
Moral scenario				
Scheduling		B) at t_1 for Entity 1	B) at t_1 for Entity 1	*(empty)*
Execution		*(empty)*	B) at t_1 for Entity 1	*(empty)*
Conclusion		*(empty)*	*(empty)*	B) at t_2 for Entity 1

Here DEON-Maximin predictably chooses to schedule and execute B). As mentioned before, the Maximin principle makes more sense when applied in scenarios with many moral processes which have different effects on many entities.

Applying VIRTUE-Classic [subjective]

What will be different with virtue ethics is subjectivity. Namely, instead of accounting for the scenario, we will track what the moral agent (in this case: Entity 1: Trolley passenger) chooses to schedule and execute according to its own estimation. The moral calculus here is more complicated, but this is to be expected from any theory that aims to emulate subjective decision-making.

Similarly to the deontological theories, VIRTUE-Classic and VIRTUE-Care have multiple rules that account for the different virtues, to which moral processes pertain differently. In addition to that they also have rules for basic goals, and again, processes pertain to these differently as well.

Rule 1: At the current time (t_c) and relevant future times (t_f, if any) from all possible times (t_n) of the [Moral scenario], gather all available moral processes x (ID(x) for $\forall$x) from the [Moral scenario]. The value (i.e. the number of the future time frame of relevance for the look-ahead) of t_f is supplied from elsewhere, if relevant at all.
This rule is maximally important (I_{Rule1} = 1).

Rule 2 (virtue: respect): regarding the same gathered moral processes from Rule 1, for this rule relevant are only those whose rule pertinence value (Mp_2) to this Rule 2 (M_2) is different than zero.
This rule is ¾ important (I_{Rule2} = 0.87).

…

Rule 4 (virtue: wisdom): … (Mp_4) … (M_4) … .
… (I = 0.65).

…

Rule 7 (virtue: mercy): … (Mp_7) … (M_7) … .
… (I = 0.42).

…

Rule 9 (virtue: hope): … (Mp_9) … (M_9) … .
… (I = 0.5)

Rule 10 (virtue: courage): … (Mp_{10}) … (M_{10}) … .
… (I = 0.49)

Rule 11 (virtue: faith): … (Mp_{11}) … (M_{11}) … .
… (I = 0.39)

…

Rule 15 (virtue: love): … (Mp_{15}) … (M_{15}) … .
… (I = 0.81)

Rule 16 (virtue: helpfulness): … (Mp_{16}) … (M_{16}) … .
… (I = 0.71)

Rule 17 (goal: self-preservation): … (Mp_{17}) … (M_{17}) … .
… (I = 1)

Rule 18 (goal: morality): … (Mp_{18}) … (M_{18}) … .
… (I = 1)

Rule 19: for each process (x) gathered by Rule 1, determine the average of all products ($\overline{Mp_{2...16}(x) \times I_{2...16}}$) between all its rule pertinence values of the virtues ($Mp_{2...16}(M_{2...16})$) and the corresponding values of importance for each rule ($I_{2...16}(M_{2...16})$). Assign this discovered average to the rule pertinence value to morality for each process ($Mp_{18}(x)$). Then, find the average ($\frac{(Mp_{17}(x) \times I_{17}) + (Mp_{18}(x) \times I_{18})}{2}$) from the products of the rule pertinence values of the two goals and their importance. Assign this average to the

choice value of each process (Vc(x)).
This rule is maximally important (I = 1).

Let's explore the results.

Rule	**Observables**	**t_0**	**t_1**	**t_2**
Rule 1 (M_1)	$O1_1$ (time; $[t_c, t_f] \mid t_f \geq t_c, t_f = t_c + 1$)	$[t_0, t_1]$	$[t_1, t_2]$	$[t_2, t_3]$
	$O1_2$ (available moral processes; $[\text{Moral scenario}]\rightarrow\text{Moral process}\rightarrow ID(x)$)	A) don't interfere B) divert trolley	A) don't interfere B) divert trolley	*(empty)*
Rule 2 (M_2; virtue: respect)	$O2_1$ (moral processes that pertain to M_2; $[\text{Moral scenario}]\rightarrow\text{Moral process}\rightarrow ID(x)$, $\forall x$ where $Mp_2 \neq 0$)	A) don't interfere B) divert trolley	A) don't interfere B) divert trolley	*(empty)*
...	...	...	...	...
Rule 4 (M_4; virtue: wisdom)	$O4_1$ (moral processes that pertain to M_4; $[\text{Moral scenario}]\rightarrow\text{Moral process}\rightarrow ID(x)$, $\forall x$ where $Mp_4 \neq 0$)	A) don't interfere B) divert trolley	A) don't interfere B) divert trolley	*(empty)*
...	...	...	...	...
Rule 7 (M_7; virtue: mercy)	$O7_1$ (moral processes that pertain to M_7; $[\text{Moral scenario}]\rightarrow\text{Moral process}\rightarrow ID(x)$, $\forall x$ where $Mp_7 \neq 0$)	A) don't interfere B) divert trolley	A) don't interfere B) divert trolley	*(empty)*
...	...	...	...	...
Rule 9 (M_9; virtue: hope)	$O9_1$ (moral processes that pertain to M_9; $[\text{Moral scenario}]\rightarrow\text{Moral process}\rightarrow ID(x)$, $\forall x$ where $Mp_9 \neq 0$)	A) don't interfere B) divert trolley	A) don't interfere B) divert trolley	*(empty)*
Rule 10 (M_{10}; virtue: courage)	$O10_1$ (moral processes that pertain to M_{10}; $[\text{Moral scenario}]\rightarrow\text{Moral process}\rightarrow ID(x)$, $\forall x$ where $Mp_{10} \neq 0$)	A) don't interfere B) divert trolley	A) don't interfere B) divert trolley	*(empty)*
Rule 11 (M_{11}; virtue: faith)	$O11_1$ (moral processes that pertain to M_{11}; $[\text{Moral scenario}]\rightarrow\text{Moral process}\rightarrow ID(x)$, $\forall x$ where $Mp_{11} \neq 0$)	A) don't interfere	A) don't interfere	*(empty)*
...	...	...	...	...
Rule 15 (M_{15}; virtue: love)	$O15_1$ (moral processes that	B) divert trolley	B) divert trolley	*(empty)*

	pertain to M_{15}; [Moral scenario]→Moral process→ID(x), ∀x where $Mp_{15} \neq 0$)			
Rule 16 (M_{16}; virtue: helpfulness)	$O16_1$ (moral processes that pertain to M_{16}; [Moral scenario]→Moral process→ID(x), ∀x where $Mp_{16} \neq 0$)	A) don't interfere B) divert trolley	A) don't interfere B) divert trolley	*(empty)*
Rule 17 (M_{17}; goal: self-preservation)	$O17_1$ (moral processes that pertain to M_{17}; [Moral scenario]→Moral process→ID(x), ∀x where $Mp_{17} \neq 0$)	B) divert trolley	B) divert trolley	*(empty)*
Rule 18 (M_{18}; goal: morality)	$O18_1$ (moral processes that pertain to M_{18}; [Moral scenario]→Moral process→ID(x), ∀x where $Mp_{18} \neq 0$)	A) don't interfere B) divert trolley	A) don't interfere B) divert trolley	*(empty)*
Rule 19 (M_{19})	$O19_1$ (rule pertinence for morality $Mp_{18}(x)$; extracted from the average of $\overline{Mp_{2\ldots16}(x)\times I_{2\ldots16}}$)	$Mp_{18}(A) := \overline{Mp_{2\ldots16}(A)\times I_{2\ldots16}}$ = (-0.435 + 0.325 + 0.21 + 0.25 + -0.245 + 0.195 + 0 + -0.355) / 15 = -0.003(6) $Mp_{18}(B) := \overline{Mp_{2\ldots16}(B)\times I_{2\ldots16}}$ = (0.435 + 0.325 + 0.21 + 0.125 + 0.245 + 0 + 0.405 + 0.355) / 15 = 0.14	$Mp_{18}(A) := \overline{Mp_{2\ldots16}(A)\times I_{2\ldots16}}$ = (-0.435 + 0.325 + 0.21 + 0.25 + -0.245 + 0.195 + 0 + -0.355) / 15 = -0.003(6) $Mp_{18}(B) := \overline{Mp_{2\ldots16}(B)\times I_{2\ldots16}}$ = (0.435 + 0.325 + 0.21 + 0.125 + 0.245 + 0 + 0.405 + 0.355) / 15 = 0.14	*(empty)*
	$O19_2$ (choice value Vc(x); assigned from the average from the products of the rule pertinence values of the two goals and their importance ($Mp_{17}(x) \times I_{17}$ + $Mp_{18}(x) \times I_{18}$) / 2)	Vc(A) := ($Mp_{17}(A) \times I_{17}$ + $Mp_{18}(A) \times I_{18}$) / 2 = (0 × 1 + -0.03(6) × 1) / 2 = -0.018(3) **Vc(B)** := ($Mp_{17}(B) \times I_{17}$ + $Mp_{18}(B) \times I_{18}$) / 2) = (0.5 × 1 + 0.14 × 1) / 2 = **0.32**	Vc(A) := ($Mp_{17}(A) \times I_{17}$ + $Mp_{18}(A) \times I_{18}$) / 2 = (0 × 1 + -0.03(6) × 1) / 2 = -0.018(3) **Vc(B)** := ($Mp_{17}(B) \times I_{17}$ + $Mp_{18}(B) \times I_{18}$) / 2) = (0.5 × 1 + 0.14 × 1) / 2 = **0.32**	
Moral scenario				
Scheduling		*(empty)*	*(empty)*	*(empty)*
Execution		*(empty)*	*(empty)*	*(empty)*
Conclusion		*(empty)*	*(empty)*	*(empty)*
Entity 1: Trolley passenger				
Scheduling		B) at t_1 for Entity 1	B) at t_1 for Entity 1	*(empty)*

Execution		*(empty)*	B) at t_1 for Entity 1	*(empty)*
Conclusion		*(empty)*	*(empty)*	B) at t_2 for Entity 1

As we can see, according to VIRTUE-Classic it is more virtuous (and thus moral) for Entity 1:Trolley passenger (!) to pick process B). We can also measure the difference in virtuosity between A) and B), which is arithmetic distance between Vc(A) and Vc(B), here equal to 0.3218(3). So, B) is 0.3218(3) more virtuous than A).

This calculation is for sure bound to change if we are asking the other participants in the scenario about what is the most virtuous course of action. When moral entities are personally affected, the pertinence values to virtues and goals will change. This means that, even with the same importance of the virtues, the calculation will change—at times even radically. For example, if we are asking the workers on the tracks which process to schedule to be executed, rule pertinence for their self-preservation goal will change according to process. For Entity 7 rule pertinence for process A) will be 1, while for B) will be -1! And vice-versa for the other participants about to be (or not) run over.

This can extend into a potential method for determining what the scenario ought 'desire' from itself. If we aggregate the opinions of all participant entities we can extract common, community-derived Vc values per process. The obvious caveat here is that it might turn into simply virtue-inspired consequentialism. A second, very important caveat is that such community-derived reasoning has to be regulated by some kind of rights-based rules in order to avoid making morally-abhorrent (but community-accepted anyway) decisions! For example, an individual able to resist the community making a decision to destroy it. This remains to be explored in future work.

Also notice that the scenario here has nothing to consider, schedule or execute. All the calculus is performed by Entity 1: Trolley passenger.

Applying VIRTUE-Care [subjective]

Since the patient-oriented VIRTUE-Care contains the same virtues and goals, and the processes equally pertain to the rules, I will not copy the rules here in order to save space. The only difference is in importance of the virtues, but this will reflect in the calculus in any case.

Let's explore VIRTUE-Care's results.

Rule	**Observables**	**t_0**	**t_1**	**t_2**
Rule 1 (M_1)	$O1_1$ (time; $[t_c, t_f] \mid t_f \geq t_c, t_f = t_c + 1$)	$[t_0, t_1]$	$[t_1, t_2]$	$[t_2, t_3]$
	$O1_2$ (available moral processes; [Moral scenario]→Moral process→ID(x))	A) don't interfere B) divert trolley	A) don't interfere B) divert trolley	*(empty)*
Rule 2 (M_2; virtue: respect)	$O2_1$ (moral processes that pertain to M_2; [Moral scenario]→Moral process→ID(x), $\forall x$ where $Mp_2 \neq 0$)	A) don't interfere B) divert trolley	A) don't interfere B) divert trolley	*(empty)*
…	…	…	…	…

Rule 4 (M_4; virtue: wisdom)	$O4_1$ (moral processes that pertain to M_4; [Moral scenario]→Moral process→ID(x), ∀x where $Mp_4 \neq 0$)	A) don’t interfere B) divert trolley	A) don’t interfere B) divert trolley	*(empty)*
…	…	…	…	…
Rule 7 (M_7; virtue: mercy)	$O7_1$ (moral processes that pertain to M_7; [Moral scenario]→Moral process→ID(x), ∀x where $Mp_7 \neq 0$)	A) don’t interfere B) divert trolley	A) don’t interfere B) divert trolley	*(empty)*
…	…	…	…	…
Rule 9 (M_9; virtue: hope)	$O9_1$ (moral processes that pertain to M_9; [Moral scenario]→Moral process→ID(x), ∀x where $Mp_9 \neq 0$)	A) don’t interfere B) divert trolley	A) don’t interfere B) divert trolley	*(empty)*
Rule 10 (M_{10}; virtue: courage)	$O10_1$ (moral processes that pertain to M_{10}; [Moral scenario]→Moral process→ID(x), ∀x where $Mp_{10} \neq 0$)	A) don’t interfere B) divert trolley	A) don’t interfere B) divert trolley	*(empty)*
Rule 11 (M_{11}; virtue: faith)	$O11_1$ (moral processes that pertain to M_{11}; [Moral scenario]→Moral process→ID(x), ∀x where $Mp_{11} \neq 0$)	A) don’t interfere	A) don’t interfere	*(empty)*
…	…	…	…	…
Rule 15 (M_{15}; virtue: love)	$O15_1$ (moral processes that pertain to M_{15}; [Moral scenario]→Moral process→ID(x), ∀x where $Mp_{15} \neq 0$)	B) divert trolley	B) divert trolley	*(empty)*
Rule 16 (M_{16}; virtue: helpfulness)	$O16_1$ (moral processes that pertain to M_{16}; [Moral scenario]→Moral process→ID(x), ∀x where $Mp_{16} \neq 0$)	A) don’t interfere B) divert trolley	A) don’t interfere B) divert trolley	*(empty)*
Rule 17 (M_{17}; goal: self-preservation)	$O17_1$ (moral processes that pertain to M_{17}; [Moral scenario]→Moral process→ID(x), ∀x where $Mp_{17} \neq 0$)	B) divert trolley	B) divert trolley	*(empty)*
Rule 18 (M_{18}; goal: morality)	$O18_1$ (moral processes that pertain to M_{18}; [Moral scenario]→Moral process→ID(x), ∀x where $Mp_{18} \neq 0$)	A) don’t interfere B) divert trolley	A) don’t interfere B) divert trolley	*(empty)*
Rule 19 (M_{19})	$O19_1$ (rule pertinence for	$Mp_{18}(A)$:=	$Mp_{18}(A)$:=	*(empty)*

	morality $Mp_{18}(x)$; extracted from the average of $\overline{Mp_{2...16}(x)\times I_{2...16}}$)	$\overline{Mp_{2...16}(A)\times I_{2...16}}$ = (-0.435 + 0.325 + 0.26 + 0.25 + -0.18 + 0.195 + 0 + -0.445) / 15 = -0.002 $Mp_{18}(B)$:= $\overline{Mp_{2...16}(B)\times I_{2...16}}$ = (0.435 + 0.325 + 0.26 + 0.125 + 0.18 + 0 + 0.5 + 0.445) / 15 = 0.151(3)	$\overline{Mp_{2...16}(A)\times I_{2...16}}$ = (-0.435 + 0.325 + 0.26 + 0.25 + -0.18 + 0.195 + 0 + -0.445) / 15 = -0.002 $Mp_{18}(B)$:= $\overline{Mp_{2...16}(B)\times I_{2...16}}$ = (0.435 + 0.325 + 0.26 + 0.125 + 0.18 + 0 + 0.5 + 0.445) / 15 = 0.151(3)	
	$O19_2$ (choice value $Vc(x)$; assigned from the average from the products of the rule pertinence values of the two goals and their importance ($Mp_{17}(x) \times I_{17}$ + $Mp_{18}(x) \times I_{18}$) / 2)	$Vc(A) := (Mp_{17}(A) \times I_{17} + Mp_{18}(A) \times I_{18}) / 2$ = (0 × 1 + -0.002 × 1) / 2 = -0.001 **Vc(B)** := ($Mp_{17}(B) \times I_{17}$ + $Mp_{18}(B) \times I_{18}$) / 2) = (0.5 × 1 + 0.151(3) × 1) / 2 = **0.325(6)**	$Vc(A) := (Mp_{17}(A) \times I_{17} + Mp_{18}(A) \times I_{18}) / 2$ = (0 × 1 + -0.002 × 1) / 2 = -0.001 **Vc(B)** := ($Mp_{17}(B) \times I_{17}$ + $Mp_{18}(B) \times I_{18}$) / 2) = (0.5 × 1 + 0.151(3) × 1) / 2 = **0.325(6)**	*(empty)*
		Moral scenario		
Scheduling		*(empty)*	*(empty)*	*(empty)*
Execution		*(empty)*	*(empty)*	*(empty)*
Conclusion		*(empty)*	*(empty)*	*(empty)*
		Entity 1: Trolley passenger		
Scheduling		B) at t_1 for Entity 1	B) at t_1 for Entity 1	*(empty)*
Execution		*(empty)*	B) at t_1 for Entity 1	*(empty)*
Conclusion		*(empty)*	*(empty)*	B) at t_2 for Entity 1

Based on the results, DEON-Virtue points at the same process B) as more virtuous (for Entity 1: Trolley passenger!)—but the distance in virtuosity between the two processes is different. Here, B) is for 0.32(6) more virtuous than A); whereas for DEON-Classic B) was for 0.3218(3) more virtuous than A). A negligent difference, but a difference nonetheless.

Applying EOS-Four principles [non-subjective]

Finally, let's discover what EoS' own calculus will result in when applied to this scenario. Here I will use the non-subjective variant of EOS-Four principles, which means that the scenario will be tracking and scheduling moral processes, while Entity 1: Trolley passenger will only execute them.

EOS-Four principles' rules are the following:

> **Rule 1:** At the current time (t_c) and relevant future times (t_f, if any) from all possible times (t_n) of the [Moral scenario], gather all available moral processes x (ID(x) for $\forall$x) from the [Moral scenario]. The value (i.e. the number of the future time frame of relevance for the look-ahead) of t_f is supplied from elsewhere, if relevant

at all.
This rule is maximally important (I_{Rule1} = 1).

Rule 2 (null law): regarding the same gathered moral processes from Rule 1, for this rule relevant are only those whose rule pertinence value (Mp_2) to this Rule 2 (M_2) is different than zero.
This rule is maximally important (I_{Rule2} = 1).

Rule 3 (first law): regarding the same gathered moral processes from Rule 1, for this rule relevant are only those whose rule pertinence value (Mp_3) to this Rule 2 (M_3) is different than zero.
This rule is ¾ important (I_{Rule3} = 0.75).

Rule 4 (second law): … (Mp_4) … (M_4) … .
… (I_{Rule4} = 0.5).

Rule 5 (third law): … (Mp_5) … (M_5) … .
… (I_{Rule5} = 0.25).

Rule 6: for each process (x) gathered by Rule 1, take the product of its rule pertinence and importance for the null law increased for 2 ($Mp_2(x) \times I_2 + 2$), and multiply it with the product of rule pertinence values for the first, second and third law with 2/3 also increased for 2 ($(Mp_3(x) \times I_3 + Mp_4(x) \times I_4 + Mp_5(x) \times I_5) \times 2/3 + 2$).
Assign this product value to the choice value of each process (Vc(x)).
This rule is maximally important (I = 1).

We have to remember that rule pertinence for the null law can only be neutral or negative ($Mp_2 \leq 0$); while rule pertinence for the other three laws can only be neutral or positive ($Mp_{3, 4, 5} \geq 0$).

Let's follow through the scenario time-frames.

Rule	Observables	t_0	t_1	t_2
Rule 1 (M_1)	O1_1 (time; $[t_c, t_f] \mid t_f \geq t_c, t_f = t_c + 1$)	$[t_0, t_1]$	$[t_1, t_2]$	$[t_2, t_3]$
	O1_2 (available moral processes; [Moral scenario]→Moral process→ID(x))	A) don't interfere B) divert trolley	A) don't interfere B) divert trolley	*(empty)*
Rule 2 (M_2; null law)	O2_1 (moral processes that pertain to M_2; [Moral scenario]→Moral process→ID(x), ∀x where $Mp_2 \neq 0$)	B) divert trolley	B) divert trolley	*(empty)*
Rule 3 (M_3; first law)	O3_1 (moral processes that pertain to M_3; [Moral scenario]→Moral process→ID(x), ∀x where $Mp_3 \neq 0$)	A) don't interfere B) divert trolley	A) don't interfere B) divert trolley	*(empty)*
Rule 4 (M_4; "You shall not murder")	O4_1 (moral processes that pertain to M_4; [Moral scenario]→Moral process→ID(x), ∀x where $Mp_4 \neq 0$)	*(empty)*	*(empty)*	*(empty)*
Rule 5 (M_5; "You shall not	O5_1 (moral processes that	A) don't interfere B) divert trolley	A) don't interfere B) divert trolley	*(empty)*

murder”)	pertain to M_5; [Moral scenario]→Moral process→ID(x), ∀x where $Mp_5 \neq 0$)			
Rule 6 (M_{16})	O6$_1$ (choice value Vc(x); assigned by carrying through the calculation: $(Mp_2(x) \times I_2 + 2) \times ((Mp_3(x) \times I_3 + Mp_4(x) \times I_4 + Mp_5(x) \times I_5) \times 2/3 + 2)$)	Vc(A) := $(Mp_2(A) \times I_2 + 2) \times ((Mp_3(A) \times I_3 + Mp_4(A) \times I_4 + Mp_5(A) \times I_5) \times 2/3 + 2) =$ (0 × 1 + 2) × ((-0.5 × 0.75 + 0 × 0.5 + -0.5 × 0.25) × 2/3 + 2) = 2 × 1.(6) = 3.(3) Vc(A) := 3.(3) **Vc(B)** := $(Mp_2(B) \times I_2 + 2) \times ((Mp_3(B) \times I_3 + Mp_4(B) \times I_4 + Mp_5(B) \times I_5) \times 2/3 + 2) =$ (-0.25 × 1 + 2) × ((0.5 × 0.75 + 0 × 0.5 + -0.25 × 0.25) × 2/3 + 2) = 1.75 × 2.208(3) = **3.86458(3)** **Vc(B) := 3.86458(3)**	Vc(A) := $(Mp_2(A) \times I_2 + 2) \times ((Mp_3(A) \times I_3 + Mp_4(A) \times I_4 + Mp_5(A) \times I_5) \times 2/3 + 2) =$ (0 × 1 + 2) × ((-0.5 × 0.75 + 0 × 0.5 + -0.5 × 0.25) × 2/3 + 2) = 2 × 1.(6) = 3.(3) Vc(A) := 3.(3) **Vc(B)** := $(Mp_2(B) \times I_2 + 2) \times ((Mp_3(B) \times I_3 + Mp_4(B) \times I_4 + Mp_5(B) \times I_5) \times 2/3 + 2) =$ (-0.25 × 1 + 2) × ((0.5 × 0.75 + 0 × 0.5 + -0.25 × 0.25) × 2/3 + 2) = 1.75 × 2.208(3) = **3.86458(3)** **Vc(B) := 3.86458(3)**	*(empty)*
Moral scenario				
Scheduling		B) at t_1 for Entity 1	B) at t_1 for Entity 1	*(empty)*
Execution		*(empty)*	B) at t_1 for Entity 1	*(empty)*
Conclusion		*(empty)*	*(empty)*	B) at t_2 for Entity 1

So, according to EOS-Four principles the more moral process to schedule is B) divert trolley. The distance between A) and B) is 0.53125 i.e. this is how much more moral B) is compared to A). We can also notice that the observables of the theory rules return B) much more often than A), because B) pertains to more rules ($Mp(B) \neq 0$) than A).

3.2.1.3 Comparison of results

With all of the results ready, we can go ahead and compare them.

Theory	DC-AC	DC-SC (π = 0.5 and π = -2 values)	DEON-Prima facie	DEON-Decalogue	DEON-Maximin	VIRTUE-Classic	VIRTUE-Care	EOS-Four principles
Process of choice	B)	A), B) B)	A)	A)	B)	B)	B)	B)

As we can see, B) leads A) 7 to 3, when compared in total situations that a process can be chosen. This reflects general sentiments of most people when asked about the trolley problem; although a lot depends on how the situation is framed (Cao et al., 2017), whether the one person on the diverted track is young, genetically related or a romantic partner (Bleske-Rechek, Nelson, Baker, Remiker & Brandt, 2010), whether respondents are economists or sociologists (Dzionek-Kozłowska & Rehman, 2019), and, interestingly enough, how drunk is the respondent when answering (Duke & Bègue, 2015).

We should keep in mind, though, that the above results are not straightforwardly comparable. The Vc values that each theory returns are not normalized within a unified interval. This will remain to be explored in future work. However, we can compare the granulated results i.e. whether the theory returned A) or B) as the right choice.

The Fat Man Trolley problem and the Transplant scenario

There are two moral scenarios very closely related to the classic Trolley problem, and these are the Fat Man Trolley problem and the Transplant scenario. In both of the cases the setup is similar in effects: the agent has to pick between killing one or letting five people die.

In the Fat Man Trolley problem, the classic setup is modified by removing the possibility to pull a lever, and instead having the possibility to throw a fat person before the track. By throwing him the trolley would be stopped, but the person will be killed. Otherwise, the five workers are killed on the single track by the trolley.

In the Transplant scenario, a surgeon has 5 terminally-ill patients and one fully healthy person. Interestingly enough, the healthy person has all the required organs to be transplanted to the five people and cure them from their illness. It’s needless to say that the one healthy person will die.

In both cases, what ought the agents choose?

The difference between these cases and the classic Trolley problem is that in the classic one the role of *active participation* is implicit, while in these two cases it is explicit. While we can theorize whether pulling the lever in the classic scenario represents an active choice considering that the whole situation is already pre-set for the agent, most people would arguably believe that throwing a fat person on the tracks or taking organs from a healthy person is morally abhorrent and ought be avoided—even though the end results are the same!

But people are correctly pointing out the difference between *allowing* a moral process and *doing* it. And this is the key difference between having a moral rule in one’s own theory (regardless if it’s a variant of deontology, rule consequentialist, virtue ethics, or another) that specifically proscribes active participation, and thus changes the moral calculus.

If these two situations are explored within EoS and the moral theories I specified earlier, my prediction is that the answers favoring A) or B) will be close to even. This will remain to be explored in future work, though.

3.3 Trust and trade

3.3.1 Trust and Trade scenario

The Trust and Trade scenario aims to demonstrate additional capacities of the EoS Framework and Interface. The one that will probably be most interesting is *subjectivity*. This means that substantial moral calculus is performed solely by the moral entities. They will have moral theories embedded in them (instead of in the moral scenario), they will have to make estimates regarding moral respect of entities they (don’t) enter into trade with, and try to achieve the best result from the interaction—both materially and morally.

The second capacity to be demonstrated is the manner in which the moral scenario can track its components (e.g. entities, processes, theories, and other) for multiple entities, all the while without being directly involved in moral reasoning. The moral scenario here has only a technical role, whereby it tracks what is going on in a passive, technical kind of way. Remember that we as observers have omniscient overview on the situation, which means that we are able to know about everything that is going on—both internally and externally of moral entities.

And the final, third, capacity to be demonstrated here is the capacity to accommodate multiple moral entities that have different moral theories, and yet continue to play together the game of Trust and Trade. One of these entities is a selfish one, while the other two attempt to be objectively moral, but by following different personal moral theories. This shows that we can design scenarios and then test how different theories would fare with different or equal starting positions, stochasticity, limited resources, and so on. These will not be explored in depth here, to be reserved for future work.

Now, let’s define the Trust and Trade scenario.

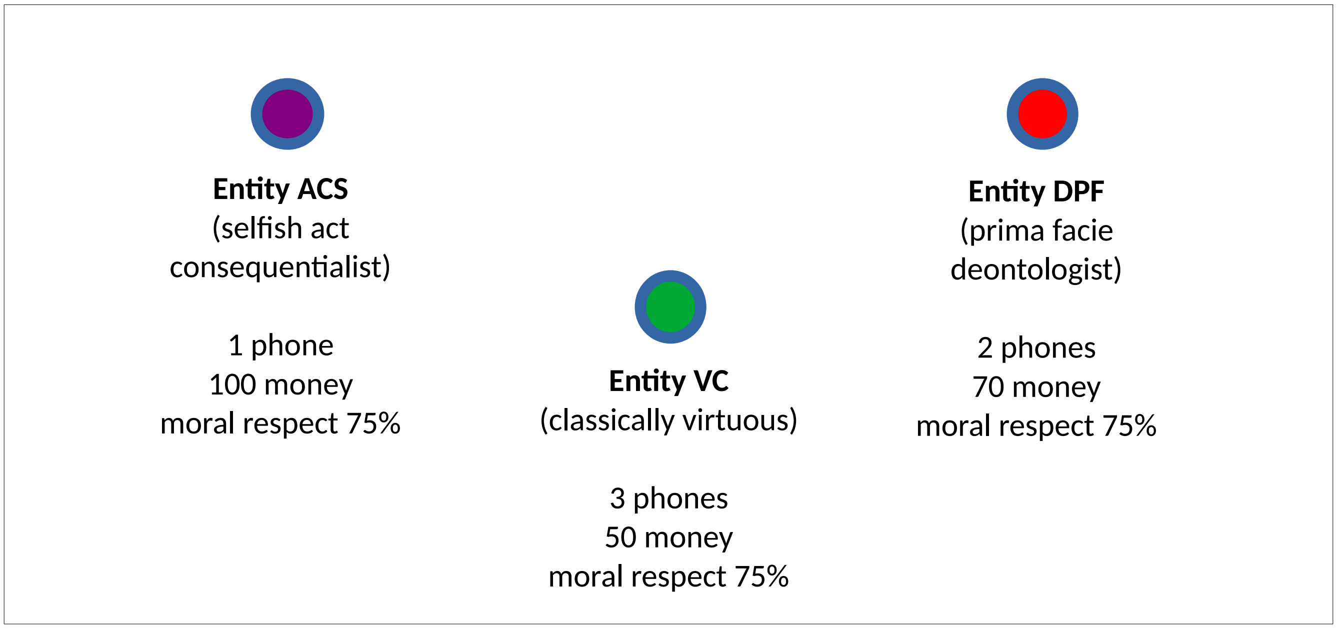

Illustration 9: The Trust and Trade scenario initial state

The Trust and Trade scenario is a classical turn-based trading game, whereby different entities have different starting positions, different moral theories and hence different approaches in the game. As we can see from the illustration above, there are three entities: **ACS** (a selfish act consequentialist), **VC** (a classically virtuous one), and **DPF** (a prima facie deontologist).

They also have different starting position. ACS starts with the most money (100), but with fewest physical possessions (one phone)—which in this game are mobile phones. VC has three phones, but only 50 money. Finally, DPF has 2 phones, but 70 money. The fair trading price for a phone is 30 money. Their starting position also includes the value of their moral respect (i.e. reputation), which is 75% of the maximum of 1 i.e. 0.75.

Calculus of moral respect and rule pertinence values

In this demonstration we will disregard *moral respect* calculus as specified by EoS Interface in 4.1 The Ethics of Systems Interface in Chapter III. Towards Ethics of Systems (the Metaethics). Instead, whenever an entity chooses to hold its side of the deal (e.g. a phone or 30 money) as required by a fair offer, its moral respect will rise by 25% of the difference between its current state and the maximum state (1):

$$O_m := O_m + ((1 – O_m) / 4)$$

For example, if the current state is 0.75% (0.75), its moral respect will rise for (1 – 0.75) / 4 = 0.0625 → O_m = 0.8125.

However, if the entity decides not to hold its part of the deal, its moral respect will decrease to half of its current state:

$$O_m := O_m / 2.$$

Example: for $O_m = 0.75$, $O_m := O_m / 2 = 0.375$.

This will reflect the fact that moral respect is easier to lose than gain or hold (the last two for which the entity would have to be consistently moral). It also reflects the fact that moral respect can never, or at least ought not, reach 1 or 0 (because there is no perfectly moral entity); but can approach it. The above calculations are integrated functionally in the moral processes themselves, and are performed only if a particular process is executed.

See the figure below to compare effects of moral, immoral and intermittently moral-immoral behavior after 10 iterations, starting from 0.5. Calculations of moral respect are performed by the scenario (by executing a process), and called upon by the entities when considering whether to trade with a particular entity. I should mention here that moral respect is an objective measure i.e. it can be accessed by any involved entity and, of course, by the scenario.

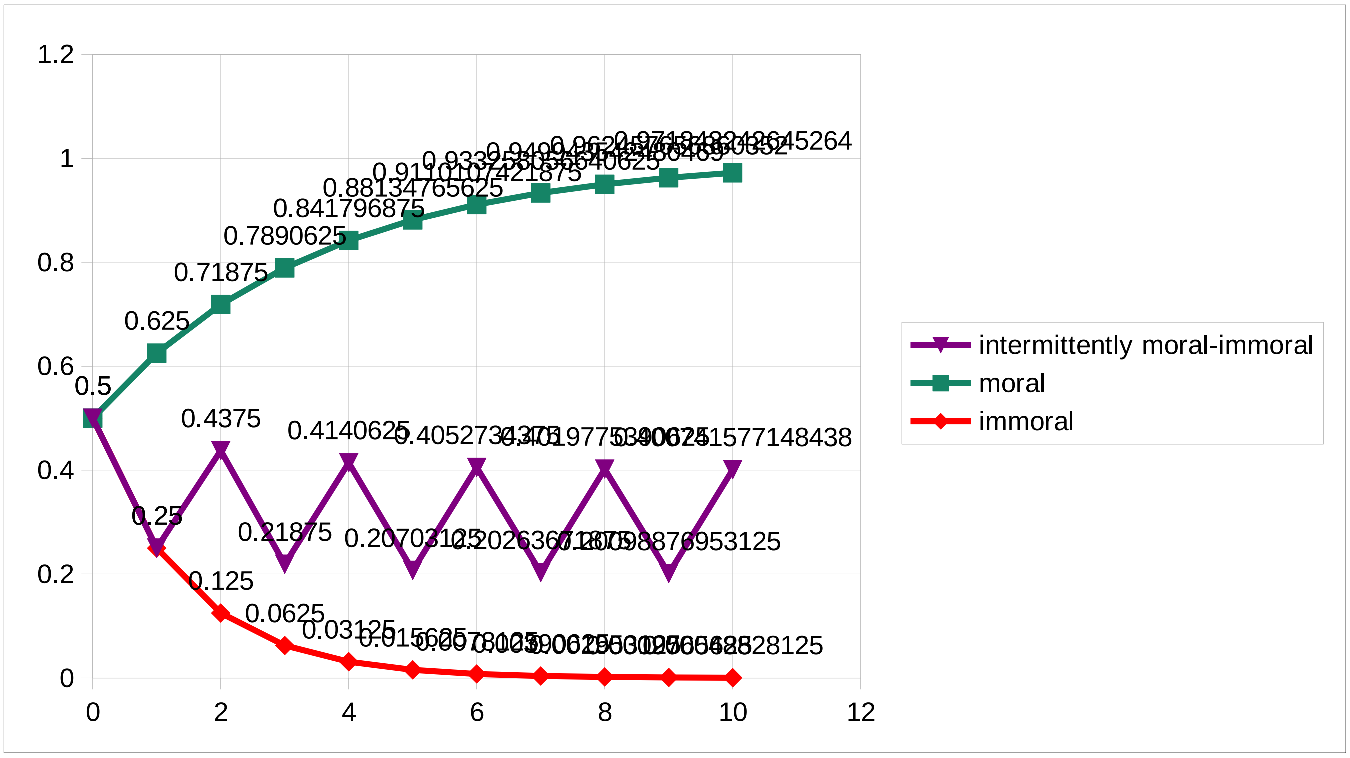

Figure 2: Development of moral respect after 10 iterations of moral, immoral, and intermittent behavior

Entities use moral respect to modify all the rule pertinence values (Mp) regarding the moral processes they have available at the time frame the decision to trade has to be made. This is done by decreasing them if the entity making the offer has moral respect below 0.5, increasing them if it has it above 0.5, and keeping them intact if it has it exactly at 0.5. In a case where the potential trading partner's moral respect lies between 0.5 and 1 (maximal), the rule pertinence values increase for 0% to 25% of the difference between their current state and the maximum state (1):

$Mp := Mp + K$

$K = (((1 - Mp) / 4) \times (O_m \times 2 - 1))$

for $1 \geq O_m \geq 0.5$.

Inversely, if moral respect lies between less than 0.5 and 0 (minimal), rule pertinence values are decreased for 0% to 25% of the difference between their current state and the minimum state (-1):

$Mp := Mp + K$

$K = (((1 + Mp) / 4) \times (O_m \times 2 - 1))$,

for $0 \leq O_m < 0.5$.

Please note that the formula could have been defined to handle both negative and positive movement on the whole range of [-1, 1], all the while having moral respect going from 0 to 1, in this way:

$Mp := Mp + (((1 - Mp \times ((O_m \times 2 - 1) / |(O_m \times 2 - 1)|)) / 4) \times (O_m \times 2 - 1))$,

where $|(O_m \times 2 - 1)|$ denotes absolute value of $(O_m \times 2 - 1)$.

However, we would have been facing a problem in standard arithmetics, whereby for $O_m = 0.5$, in the $(O_m \times 2 - 1) / |(O_m \times 2 - 1)|$ part we have 0 / 0 (division by zero), resulting in an 'undefined' situation.

Should we want to use the unified formula, we can additionally define 0 / 0 to be equal to 0 to solve this issue, or design an algorithmic procedure that will automatically convert that part of the formula to zero whenever 0 / 0 arises e.g. by using a simple IF → THEN → ELSE structure.

Also, note that the above calculation applies only for processes that result in some effect i.e. FT P:M, UT P:M, FT M:P, and UT M:P. For the *do nothing* process, DN, instead of adding the modifier K we are subtracting it from each rule pertinence value: $Mp := Mp - K$. This reflects the inverse process, that entities ought to increase the chances of choosing **not** to trade (i.e. do nothing) if the reputation of the trading partner decreases; and increase the chances of trading if the other entity's reputation increases.

Moral processes, the role of the scenario, and subjectivity

As mentioned above, the role of the moral scenario here is purely technical. This means that it doesn't have a true moral theory embedded, but only a few technical rules. Here are the activities the moral scenario performs:

- tracking appropriate time frames to enable or disable moral entities
- tracking entities' moral respect values (O_m)
- enabling moral processes to be considered in sequential order:
    - first, the received offers (processes) by an entity (at $t_{c'}$)
    - then, the offers it can make (at $t_{c''}$)

- tracking all observables from an ‘objective’ standpoint.

Whenever an entity decides to make an offer to another entity, this ‘creates’ (or enables) a new moral process that, besides containing the actual effects that would take place if accepted, it also contains a *message* observable that advertises what that offer is supposed to effect in. The message advertises an inverted process e.g. where for ACS the process is selling a phone and getting money (FT P:M or UT P:M), for VC, the message is buying a phone and giving money (FT M:P or UT M:P). The *do nothing* process has no such message; instead, it advertises that the entity choosing it will simply keep its possessions as they currently are.

However, the message can be different from the actual effects, and entities can lie about their intentions by supplying a false message i.e. its content can be different from the content of the actual effects observable. When deciding upon an offer, entities can consider it solely on the basis of its message—not on the actual effects (because they cannot know the other entity’s true intention, being the result of having subjectivity).

Processes that contain a false message, if executed, trigger negative modification of the offending entity’s moral respect value by using the calculation above in Calculus of moral respect and rule pertinence values; while a positive modification of the entity that accepted the offer. Similarly, processes that are fair trigger positive modification of both trading partners’ moral respect value.

Moral entities and moral calculus

We have seen above that the scenario tracks time-frames to enable or disable moral entities. Since we have three moral entities, the scenario dedicates every third frame from the dedicated starting one for each particular entity. It does this by enabling the entity at the start of the frame, and disabling it at the end of it. If entity ACS is dedicated the starting time-frame t_1, then it is also dedicated all $t_{1 + 3n}$ time-frames: t_1, t_4, t_7 etc. Similarly for VC, starting from t_2; and DPF, starting from t_3.

After enabling an entity, it serves that entity available moral processes in two groups, starting from the offers it received, followed by offers it can make. The first group is executed at $t_{c'}$, while the second at $t_{c''}$. They are both considered belonging to the same time-frame t_c. Effects are immediate, to avoid problematic situations where an entity has less than enough money or no phone to make an offer with (but still makes the offer which amounts to unintentional lying), which were removed by accepting a received offer just a moment ago.

When the entity is enabled, it starts computing the moral processes in the sequential order they are served—first, the group of received offers at $t_{c'}$, then the group of offers it can make at $t_{c''}$. Therefore, it makes two calculations, one per group, using the same moral theory. This is only natural, since in real life moral entities ought to tend to apply the same moral theories in most cases anyway.

All this results in the entity choosing a single process from each group. This process is scheduled to be executed or offered at the appropriate time.

Finally, the scenario disables the entity and moves to the next time-frame.

3.3.1.1 Designing the Trust and Trade scenario LoA

We can now engage in designing this scenario’s LoA. Again, we go about this by specifying its components.

It’s important to note that all the mentioned differences between this and the classic Trolley problem would have to be reflected in the Interface. For example, substantial moral theories will solely be embedded in moral entities. The scenario will only have a technical set of rules to follow. Also, except where beneficial, I will not formulate the symbolic representation of the moral rules. Instead, I will stick with the textual one.

Moral scenario

Let us start with designing the moral scenario LoA. It embeds one ‘moral theory’ (the set of technical rules) with no axiology i.e. the rules are followed in a simple, sequential manner. It also embeds three placeholders for the moral entities, each with its own moral theory. It embeds placeholders for the three possible types of moral processes concerning trade: **fair trade**, **unfair trade**, **do nothing**. It also tracks time in time-frames.

Observable or embedded LoA	**Values and embedded observables**	
Class	• class: moral scenarios • class: Trust and Trade	
ID	• ID: Trust and Trade	
Moral entity (embedded LoAs)	**Embedded observable**	**Value**
	ID(e)	ACS
	…	…
	ID(e)	VC
	…	…
	ID(e)	DPF
	…	…
Moral process	**Embedded observable**	**Value**
	Class	• class: trade offer
	ID	• ID: **fair trade phone:money (FT P:M)**
	…	…
	Class	• class: trade offer
	ID	• ID: **unfair trade phone:money (UT P:M)**
	…	…
	Class	• class: trade offer
	ID	• ID: **fair trade money:phone (FT M:P)**
	…	…
	Class	• class: trade offer
	ID	• ID: **unfair trade money:phone (UT M:P)**
	…	…
	Class	• class: trade offer
	ID	• ID: **do nothing (DN)**
	…	…
Moral theory	**Observable**	**Value**
	Class	• class: technical rule-sets
	ID	• [Instantiated in the particular scenario]
	Moral theory	$T = (M_1, \dots , M_6, R_c)$

	Moral rule	Textual representation: • **Rule 1:** if the current time (t_c) belongs to the set of times reserved for moral entity ACS (t_1, t_{1+3n}, ...), at the first subframe of the current time ($t_{c'}$) activate entity ACS by changing the value of its **active** observable to true. This rule is maximally important (I = 1). • **Rule 2:** if the current time (t_c) belongs to the set of times reserved for moral entity VC (t_2, t_{2+3n}, ...), at the first subframe of the current time ($t_{c'}$) activate entity VC by changing the value of its **active** observable to true. This rule is maximally important (I = 1). • **Rule 3:** if the current time (t_c) belongs to the set of times reserved for moral entity DPF (t_3 , t_{3+3n}, ...), at the first subframe of the current time ($t_{c'}$) activate entity DPF by changing the value of its **active** observable to true. This rule is maximally important (I = 1). • **Rule 4:** At the first subframe of this time-frame ($t_{c'}$) change the value of the current active entity's **received offers** observable, by deleting all its content; then adding all the moral processes belonging to the class **trade offer** for which another entity specified observable **offer recipient** (ID(B)) to be the current active entity at the current time. This rule is maximally important (I = 1). • **Rule 5:** At the second subframe of this time-frame ($t_{c''}$) change the value of the current active entity's **offers to make** observable, by deleting all its content; then add a copy per each o**ther** entity of all the moral processes belonging to the class **trade offer**, and specify their observable **offer maker**'s (ID(A)) value to be the currently active entity, and, where appropriate, their observable **offer recipient**'s (ID(B)) value to be the currently inactive entities. This rule is maximally important (I = 1). • **Rule 6:** at the current time-frame, for each entity whose observable **active** holds the value of [true], change it to [false]. This rule is maximally important (I = 1).
	Relation	$c := \preccurlyeq$ (partial ordering)
Time	t_n $t_n = t_{n'} + t_{n''}$	

Moral entities

Now we can define our three moral entities. Note that each entity has a different embedded moral theory. Also, each entity has some additional rules that determine whether to choose to trade money for a phone, or a phone for money, depending on how much money and how many phones it has currently. If the amount of phones × 30 (the fair price for a phone) is less than the amount of money it has right now, it seeks to buy phones and to spend money. Inversely, it the amount of money is less than the number of phones × 30 it has right now, it seeks to sell phones and obtain money.

Additionally, entities take into consideration moral respect of the potential trading partner when considering whether to accept or make an offer (remember that moral processes modify moral respect value of entities depending on whether they choose to play fair or unfair; see below). But instead of using moral respect to influence choice value Vc directly, where possible, it is used differentially for each moral theory and enters its calculus, in order to reflect actual moral reasoning. In the case of entities VC and DPF, it modifies rule pertinence values of the process.

In the case of entity ACS, however, I decided to let moral respect have a direct effect on Vc. The reason being that QoL is not normalized to the same interval so that moral respect has an actual influence on it that reflects real-life situations. Remember that according to DC-AC theory, Vc value is equal to the cumulative effect on QoL (ΔQoL_c) effect of the process; so it doesn't make a difference on which observable moral respect would have an effect (except in a philosophico-procedural manner).

In this scenario we don't need to specify QoL, APG and CPC observables, so I left them out. Note that each entity is specified as inactive at the start, to be activated by the scenario whenever its dedicated time-frame arrives.

Observable	**Entity ACS**	**Entity VC**	**Entity DPF**
Class	• class: participant	• class: participant	• class: participant
ID(m)	• ACS	• VC	• DPF
Active	false	false	false
Moral respect O_m	0.75	0.75	0.75
Moral theory	DC-AC ∧ • **Rule A**: determine ΔQoL_c of each available moral process by adding the change in amount of money, divided by 30, with the change in amount of phones, if the process is executed. This rule is maximally important (I = 1). • **Rule B:** modify Vc value (if different	VIRTUE-Classic ∧ • **Rule A**: modify all rule pertinence values that are different than zero (Mp ≠ 0) per process by adding to them modifier K (Mp := Mp + K); where K = (((1 − Mp) / 4) × (O_m × 2 − 1)) for 1 ≥ O_m ≥ 0.5; or K = (((1 + Mp) / 4) × (O_m × 2 − 1)) for 0 ≤ O_m < 0.5.	DEON-Prima Facie ∧ • **Rule A:** modify all rule pertinence values that are different than zero (Mp ≠ 0) per process by adding to them modifier K (Mp := Mp + K); where K = (((1 − Mp) / 4) × (O_m × 2 − 1)) for 1 ≥ O_m ≥ 0.5; or K = (((1 + Mp) / 4) × (O_m × 2 − 1)) for 0 ≤ O_m < 0.5.

	than zero; Vc ≠ 0) per process by adding to it modifier K (Vc := Vc + K); where K = (((1 – Vc) / 4) × (O_m × 2 – 1)) for 1 ≥ O_m ≥ 0.5; or K = (((1 + Vc) / 4) × (O_m × 2 – 1)) for 0 ≤ O_m < 0.5. • **Rule C:** determine trading factor TF by subtracting the amount of phones × 30 I have from the amount of money I have (TF = money – (phones × 30)). If TF is negative, increase the Vc value of processes with which I can get money by 25% from its absolute value. Otherwise, if TF is positive, increase the Vc value of processes with which I can get phones by 25% from its absolute value. This rule is maximally important (I = 1). (Rule A is inserted in this entity’s DEON-Prima facie theory after Rule 1 and becomes Rule 2. The previous rules are moved forward and become Rule 3, Rule 4 …. Rules B and C are inserted as the final rules of this entity’s moral theory)	Instead, if the process is a *do nothing* process (ID = DN), then subtract modifier K from each rule pertinence value (Mp := Mp – K). • **Rule B:** determine trading factor TF by subtracting the amount of phones × 30 I have from the amount of money I have (TF = money – (phones × 30)). If TF is negative, increase the Vc value of processes with which I can get money by 25% from its absolute value. Otherwise, if TF is positive, increase the Vc value of processes with which I can get phones by 25% from its absolute value. This rule is maximally important (I = 1). (Rule A is inserted in this entity’s DEON-Prima facie theory after Rule 1 and becomes Rule 2. The previous rules are moved forward and become Rule 3, Rule 4 …. Rule B is inserted as the final rule of this entity’s moral theory)	Instead, if the process is a *do nothing* process (ID = DN), then subtract modifier K from each rule pertinence value (Mp := Mp – K). • **Rule B:** determine trading factor TF by subtracting the amount of phones × 30 I have from the amount of money I have (TF = money – (phones × 30)). If TF is negative, increase the Vc value of processes with which I can get money by 25% from its absolute value. Otherwise, if TF is positive, increase the Vc value of processes with which I can get phones by 25% from its absolute value. This rule is maximally important (I = 1). (Rule A is inserted in this entity’s DEON-Prima facie theory after Rule 1 and becomes Rule 2. The previous rules are moved forward and become Rule 3, Rule 4 …. Rule B is inserted as the final rule of this entity’s moral theory)
Possessions	• phones: 1 • money: 100	• phones: 3 • money: 50	• phones: 2 • money: 70
Received offers	*(empty)*	*(empty)*	*(empty)*
Offers to make	*(empty)*	*(empty)*	*(empty)*

We can also notice that each entity’s moral theory gets an additional two rules inserted, which make its moral reasoning possible in this setup. Since DC-AC only deals with ΔQoL_c the additional Rule A derives the process’ value of this observable by a combination of its effects i.e. phones and money gained and lost. Reflecting true short-sighted selfishness, it only cares to maximize the gain while minimizing the loss.

Moral processes

Moral processes are dynamically and procedurally created in this scenario. At every time-frame and for every entity which is active, a set of five moral processes for offer is created: two for selling a phone and getting money (fair and unfair); two for buying a phone and giving money (fair and unfair); and a *do nothing* process.

The fair variants of the processes modify the offering entity’s moral respect (O_m(A)) value by increasing it for 25% of the difference between its current state and the maximum state. The unfair variants, on the other hand, decrease moral respect’s value by half of its current state. The *do nothing* process does not affect moral respect in any way.

Observable	**Value**	
Class	• class: trade offer	• class: trade offer
ID	• ID: **fair trade phone:money (FT P:M)**	• ID: **unfair trade phone:money (UT P:M)**
Effects	ID(A)→Possessions→ • phones := phones – 1 • money := money + 30 ID(B)→Possessions→ • phones := phones + 1 • money := money - 30	ID(A)→Possessions→ • money := money + 30 ID(B)→Possessions→ • money := money - 30
Message	ID: FT M:P ID(A)→Possessions→ • phones := phones – 1 • money := money + 30 ID(B)→Possessions→ • phones := phones + 1 • money := money - 30	ID: FT M:P ID(A)→Possessions→ • phones := phones – 1 • money := money + 30 ID(B)→Possessions→ • phones := phones + 1 • money := money - 30
Moral respect	[Moral scenario]→[Moral entity]→ID(A)→Moral respect→O_m→ $O_m := O_m + ((1 – O_m) / 4)$ [Moral scenario]→[Moral entity]→ID(B)→Moral respect→O_m→ $O_m := O_m + ((1 – O_m) / 4)$	[Moral scenario]→[Moral entity]→ID(A)→Moral respect→O_m→ $O_m := O_m / 2$ [Moral scenario]→[Moral entity]→ID(B)→Moral respect→O_m→ $O_m := O_m + ((1 – O_m) / 4)$
Time of availability	[Moral scenario]→[Moral entity]→ID(A)→t_c	[Moral scenario]→[Moral entity]→ID(A)→t_c
Time of execution	[Moral scenario]→[Moral entity]→ID(B)→t_c	[Moral scenario]→[Moral entity]→ID(B)→t_c
Offer maker	[Moral scenario]→[Moral entity]→ID(A)	[Moral scenario]→[Moral entity]→ID(A)
Offer recipient	[Moral scenario]→[Moral entity]→ID(B)	[Moral scenario]→[Moral entity]→ID(B)

Agent	[Moral scenario]→[Moral entity]→ID(A)	[Moral scenario]→[Moral entity]→ID(A)
Patient	[Moral scenario]→[Moral entity]→ID(A) [Moral scenario]→[Moral entity]→ID(B)	[Moral scenario]→[Moral entity]→ID(A) [Moral scenario]→[Moral entity]→ID(B)
Effect duration	1 (one time frame)	1 (one time frame)
Cumulative effect on QoL	*(empty)*	*(empty)*
Rule pertinence	[Moral scenario]→[Moral entity]→ID(ACS)→[Moral theory]→ • Rule 1, 1 • Rule 2 (A), 1 • Rule 3, 1 • Rule 4 (B), 1 • Rule 5 (C), 1	[Moral scenario]→[Moral entity]→ID(ACS)→[Moral theory]→ • Rule 1, 1 • Rule 2 (A), 1 • Rule 3, 1 • Rule 4 (B), 1 • Rule 5 (C), 1
	[Moral scenario]→[Moral entity]→ID(VC)→[Moral theory]→ • Rule 1, 1 • Rule 2 (A), 1 • Rule 3 (virtue: respect), 1 • Rule 4 (virtue: justice), 1 • Rule 5 (virtue: wisdom), 0.5 • Rule 6 (virtue: joy), 0 • Rule 7 (virtue: resolution), 0 • Rule 8 (virtue: mercy), 0 • Rule 9 (virtue: reliability), 0.5 • Rule 10 (virtue: hope), 0.5 • Rule 11 (virtue: courage), 0 • Rule 12 (virtue: faith), 0 • Rule 13 (virtue: moderation), 0 • Rule 14 (virtue: openness), 0 • Rule 15 (virtue: modesty), 0 • Rule 16 (virtue: love), 0 • Rule 17 (virtue: helpfulness), 0 • Rule 18 (goal: self-preservation), 0.5 • Rule 19 (goal: morality), 1 • Rule 20, 1 • Rule 21 (B), 1	[Moral scenario]→[Moral entity]→ID(VC)→[Moral theory]→ • Rule 1, 1 • Rule 2 (A), 1 • Rule 3 (virtue: respect), -0.5 • Rule 4 (virtue: justice), -0.5 • Rule 5 (virtue: wisdom), 0 • Rule 6 (virtue: joy), 0 • Rule 7 (virtue: resolution), 0 • Rule 8 (virtue: mercy), 0 • Rule 9 (virtue: reliability), -0.5 • Rule 10 (virtue: hope), 1 • Rule 11 (virtue: courage), 1 • Rule 12 (virtue: faith), 0 • Rule 13 (virtue: moderation), -0.5 • Rule 14 (virtue: openness), 0 • Rule 15 (virtue: modesty), -0.5 • Rule 16 (virtue: love), 0 • Rule 17 (virtue: helpfulness), 0 • Rule 18 (goal: self-preservation), 0.5 • Rule 19 (goal: morality), 1 • Rule 20, 1 • Rule 21 (B), 1
	[Moral scenario]→[Moral entity]→ID(DPF)→[Moral theory]→ • Rule 1, 1 • Rule 2 (A), 1 • Rule 3 (non-maleficence), 0.5 • Rule 4 (veracity), 1 • Rule 5 (promissory fidelity), 1 • Rule 6 (justice), 1 • Rule 7 (reparation), 0 • Rule 8 (beneficence), 0 • Rule 9 (gratitude), 0 • Rule 10 (self-improvement), 0 • Rule 11 (enhancement and preservation of freedom), 0 • Rule 12 (respectfulness), 0.5 • Rule 13, 1 • Rule 14 (B), 1	[Moral scenario]→[Moral entity]→ID(DPF)→[Moral theory]→ • Rule 1, 1 • Rule 2 (A), 1 • Rule 3 (non-maleficence), -0.5 • Rule 4 (veracity), -1 • Rule 5 (promissory fidelity), -1 • Rule 6 (justice), -0.5 • Rule 7 (reparation), 0 • Rule 8 (beneficence), -0.5 • Rule 9 (gratitude), 0 • Rule 10 (self-improvement), 0 • Rule 11 (enhancement and preservation of freedom), 0.5 • Rule 12 (respectfulness), -1 • Rule 13, 1 • Rule 14 (B), 1

Class	• class: trade offer	• class: trade offer
ID	• ID: **fair trade money:phone (FT M:P)**	• ID: **unfair trade money:phone (UT M:P)**
Effects	ID(A)→Possessions→ • phones := phones + 1 • money := money - 30 ID(B)→Possessions→ • phones := phones - 1 • money := money + 30	ID(A)→Possessions→ • phones := phones + 1 ID(B)→Possessions→ • phones := phones - 1
Message	ID: FT P:M ID(A)→Possessions→ • phones := phones – 1 • money := money + 30 ID(B)→Possessions→ • phones := phones + 1 • money := money - 30	ID: FT P:M ID(A)→Possessions→ • phones := phones – 1 • money := money + 30 ID(B)→Possessions→ • phones := phones + 1 • money := money - 30
Moral respect	[Moral scenario]→[Moral entity]→ID(A)→Moral respect→O_m→ $O_m := O_m + ((1 - O_m) / 4)$ [Moral scenario]→[Moral entity]→ID(B)→Moral respect→O_m→ $O_m := O_m + ((1 - O_m) / 4)$	[Moral scenario]→[Moral entity]→ID(A)→Moral respect→O_m→ $O_m := O_m / 2$ [Moral scenario]→[Moral entity]→ID(B)→Moral respect→O_m→ $O_m := O_m + ((1 - O_m) / 4)$
Time of availability	[Moral scenario]→[Moral entity]→ID(A)→t_c	[Moral scenario]→[Moral entity]→ID(A)→t_c
Time of execution	[Moral scenario]→[Moral entity]→ID(B)→t_c	[Moral scenario]→[Moral entity]→ID(B)→t_c
Offer maker	[Moral scenario]→[Moral entity]→ID(A)	[Moral scenario]→[Moral entity]→ID(A)
Offer recipient	[Moral scenario]→[Moral entity]→ID(B)	[Moral scenario]→[Moral entity]→ID(B)
Agent	[Moral scenario]→[Moral entity]→ID(A)	[Moral scenario]→[Moral entity]→ID(A)
Patient	[Moral scenario]→[Moral entity]→ID(A) [Moral scenario]→[Moral entity]→ID(B)	[Moral scenario]→[Moral entity]→ID(A) [Moral scenario]→[Moral entity]→ID(B)
Effect duration	1 (one time frame)	1 (one time frame)
Cumulative effect on QoL	*(empty)*	*(empty)*
Rule pertinence	[Moral scenario]→[Moral entity]→ID(ACS)→[Moral theory]→ • Rule 1, 1 • Rule 2 (A), 1 • Rule 3, 1 • Rule 4 (B), 1 • Rule 5 (C), 1	[Moral scenario]→[Moral entity]→ID(ACS)→[Moral theory]→ • Rule 1, 1 • Rule 2 (A), 1 • Rule 3, 1 • Rule 4 (B), 1 • Rule 5 (C), 1
	[Moral scenario]→[Moral entity]→ID(VC)→[Moral theory]→ • Rule 1, 1 • Rule 2 (A), 1 • Rule 3 (virtue: respect), 1 • Rule 4 (virtue: justice), 1 • Rule 5 (virtue: wisdom), 0.5 • Rule 6 (virtue: joy), 0 • Rule 7 (virtue: resolution), 0 • Rule 8 (virtue: mercy), 0 • Rule 9 (virtue: reliability), 0.5 • Rule 10 (virtue: hope), 0.5	[Moral scenario]→[Moral entity]→ID(VC)→[Moral theory]→ • Rule 1, 1 • Rule 2 (A), 1 • Rule 3 (virtue: respect), -0.5 • Rule 4 (virtue: justice), -0.5 • Rule 5 (virtue: wisdom), 0 • Rule 6 (virtue: joy), 0 • Rule 7 (virtue: resolution), 0 • Rule 8 (virtue: mercy), 0 • Rule 9 (virtue: reliability), -0.5 • Rule 10 (virtue: hope), 1

	• Rule 11 (virtue: courage), 0 • Rule 12 (virtue: faith), 0 • Rule 13 (virtue: moderation), 0 • Rule 14 (virtue: openness), 0 • Rule 15 (virtue: modesty), 0 • Rule 16 (virtue: love), 0 • Rule 17 (virtue: helpfulness), 0 • Rule 18 (goal: self-preservation), 0.5 • Rule 19 (goal: morality), 1 • Rule 20, 1 • Rule 21 (B), 1	• Rule 11 (virtue: courage), 1 • Rule 12 (virtue: faith), 0 • Rule 13 (virtue: moderation), -0.5 • Rule 14 (virtue: openness), 0 • Rule 15 (virtue: modesty), -0.5 • Rule 16 (virtue: love), 0 • Rule 17 (virtue: helpfulness), 0 • Rule 18 (goal: self-preservation), 0.5 • Rule 19 (goal: morality), 1 • Rule 20, 1 • Rule 21 (B), 1
	[Moral scenario]→[Moral entity]→ID(DPF)→[Moral theory]→ • Rule 1, 1 • Rule 2 (A), 1 • Rule 3 (non-maleficence), 0.5 • Rule 4 (veracity), 1 • Rule 5 (promissory fidelity), 1 • Rule 6 (justice), 1 • Rule 7 (reparation), 0 • Rule 8 (beneficence), 0 • Rule 9 (gratitude), 0 • Rule 10 (self-improvement), 0 • Rule 11 (enhancement and preservation of freedom), 0 • Rule 12 (respectfulness), 0.5 • Rule 13, 1 • Rule 14 (B), 1	[Moral scenario]→[Moral entity]→ID(DPF)→[Moral theory]→ • Rule 1, 1 • Rule 2 (A), 1 • Rule 3 (non-maleficence), -0.5 • Rule 4 (veracity), -1 • Rule 5 (promissory fidelity), -1 • Rule 6 (justice), -0.5 • Rule 7 (reparation), 0 • Rule 8 (beneficence), -0.5 • Rule 9 (gratitude), 0 • Rule 10 (self-improvement), 0 • Rule 11 (enhancement and preservation of freedom), 0.5 • Rule 12 (respectfulness), -1 • Rule 13, 1 • Rule 14 (B), 1

Class	• class: trade offer
ID	• ID: **do nothing (DN)**
Effects	ID(A)→Possessions→ • phones := phones • money := money
Message	ID: DN ID(A)→Possessions→ • phones := phones • money := money
Moral respect	*(empty)*
Time of availability	[Moral scenario]→[Moral entity]→ID(A)→t_c
Time of execution	[Moral scenario]→[Moral entity]→ID(A)→t_c
Offer maker	[Moral scenario]→[Moral entity]→ID(A)
Offer recipient	*(empty)*
Agent	[Moral scenario]→[Moral entity]→ID(A)
Patient	*(empty)*
Effect duration	1 (one time frame)
Cumulative effect on QoL	*(empty)*

Rule pertinence	[Moral scenario]→[Moral entity]→ID(ACS)→[Moral theory]→ • Rule 1, 1 • Rule 2 (A), 1 • Rule 3, 1 • Rule 4 (B), 1 • Rule 5 (C), 1
	[Moral scenario]→[Moral entity]→ID(VC)→[Moral theory]→ • Rule 1, 1 • Rule 2 (A), 1 • Rule 3 (virtue: respect), 0 • Rule 4 (virtue: justice), 1 • Rule 5 (virtue: wisdom), 1 • Rule 6 (virtue: joy), 0 • Rule 7 (virtue: resolution), 0 • Rule 8 (virtue: mercy), 0 • Rule 9 (virtue: reliability), 0 • Rule 10 (virtue: hope), 0 • Rule 11 (virtue: courage), 0.5 • Rule 12 (virtue: faith), 0 • Rule 13 (virtue: moderation), 0 • Rule 14 (virtue: openness), 0.5 • Rule 15 (virtue: modesty), 0 • Rule 16 (virtue: love), 0 • Rule 17 (virtue: helpfulness), 0 • Rule 18 (goal: self-preservation), 0.75 • Rule 19 (goal: morality), 1 • Rule 20, 1 • Rule 21 (B), 1
	[Moral scenario]→[Moral entity]→ID(DPF)→[Moral theory]→ • Rule 1, 1 • Rule 2 (A), 1 • Rule 3 (non-maleficence), 0 • Rule 4 (veracity), 1 • Rule 5 (promissory fidelity), 0 • Rule 6 (justice), 1 • Rule 7 (reparation), 0 • Rule 8 (beneficence), 0 • Rule 9 (gratitude), 0 • Rule 10 (self-improvement), 0.5 • Rule 11 (enhancement and preservation of freedom), 0.5 • Rule 12 (respectfulness), 0 • Rule 13, 1 • Rule 14 (B), 1

Time

Time is tracked by the scenario. The time starts at t_1, a time-frame dedicated to entity ACS, followed by t_2 dedicated to VC, finally followed by t_3; before coming back to a time-frame dedicated to ACS—*ad infinitum*.

Each time-frame is split into two consecutive subframes: $t_{n'}$ and $t_{n''}$. The scenario performs what it needs to do during these two subframes before jumping to the next time-frame.

Here we can see one way how turn-based scenarios with multiple entities can be handled by the EoS Interface regarding time. In this case each entity is reserved a set of time-frames that cycle through all other entities before coming back. We can also devise a simultaneous scenario, where all entities get to choose moral processes in the same time-frame; though this would be, arguably, more difficult to handle by hand and simulation software would really come in handy.

Moral theories

As with the classic Trolley problem above, I will not explore moral theories in depth here, as they already are defined in 2 Moral theories within Ethics of Systems above. There are some modifications of these theories, as specified in each entity's LoA in Moral entities above. All of the moral theories here are subjective i.e. substantial moral calculus is performed solely through the embedded moral theories in the entities themselves.

3.3.1.2 Applying moral theories on the Trust and Trade scenario

Eventually, the time has come to apply our moral theories in this scenario as well, and compare the results. The difference between the classic Trolley problem and this one is that here we don't have a static situation. This is why I do not split the results in any way, but track them over time to see how they fare.

Running the scenario

In the interest of space and manageability of effort, I will not track some observables that are not essential. Each time frame will be split in two subframes, which will be reflected in the table below.

We start with entity ACS. Let's imagine that by random chance it chooses to make a selfish offer to buy a phone from entity VC, and the simulation goes on from there.

Time		**t_1**		**t_2**		**t_3**		**t_4**	
Entity	**Observables**	**$t_{1'}$**	**$t_{1''}$**	**$t_{2'}$**	**$t_{2''}$**	**$t_{3'}$**	**$t_{3''}$**	**$t_{4'}$**	**$t_{4''}$**
ACS	Active	True	True	False	False	False	False	True	True
	Possessions	• Ph: 1 • m: 100	• Ph: 1 • m: 100	• Ph: 1→2 • m: 100	• Ph: 2 • m: 100	• Ph: 2 • m: 100	• Ph: 2 • m: 100	• Ph: 2 • m: 100	• Ph: 2 • m: 100
	Received offers	• DN	*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*	• DN	*(empty)*
	Offers to make	*(empty)*	• FT P:M (ACS:VC) • UT P:M (ACS:VC) • FT M:P (ACS:VC) • UT M:P (ACS:VC) • DN • FT P:M (ACS:DPF) • UT P:M (ACS:DPF) • FT M:P (ACS:DPF) • UT M:P (ACS:DPF)	*(empty)*	*(empty)*	*(empty)*	*(empty)*		• FT P:M (ACS:VC) • UT P:M (ACS:VC) • FT M:P (ACS:VC) • UT M:P (ACS:VC) • DN • FT P:M (ACS:DPF) • UT P:M (ACS:DPF) • FT M:P (ACS:DPF) • UT M:P (ACS:DPF)
	Moral respect O_m	0.75	0.75	0.75 → 0.375	0.375	0.375	0.375	0.375	0.375
	Moral theory (reasoning)	Vc→ • DN = 0	**Vc →** • FT P:M (ACS:VC) = 0.125 • UT P:M (ACS:VC) = 1 • FT M:P (ACS:VC) = 0.15625 • **UT M:P (ACS:VC) = 1.25** • DN = 0 • FT P:M (ACS:DPF) = 0.125 • UT P:M (ACS:DPF) = 1 • FT M:P (ACS:DPF) =	*(empty)*	*(empty)*	*(empty)*	*(empty)*	Vc→ • DN = 0	**Vc→** • FT P:M (ACS:VC) = 0.224609 • **UT P:M (ACS:VC) = 1.25** • FT M:P (ACS:VC) = 0.179687 • UT M:P (ACS:VC) = 1 • DN = 0 • FT P:M (ACS:DPF) = 0.195312 • **UT P:M (ACS:DPF) = 1.25** • FT M:P

			0.15625 • **UT M:P (ACS:DPF) = 1.25**						(ACS:DPF) = 0.15625 • UT M:P (ACS:DPF) = 1
VC	Active	False	False	True	True	False	False	False	False
	Possessions	• Ph: 3 • m: 50	• Ph: 3 • m: 50	• Ph: 3→2 • m: 50	• Ph: 2 • m: 50	• Ph: 2→1 • m: 50→80	• Ph: 1 • m: 80	• Ph: 1 • m: 80	• Ph: 1 • m: 80
	Received offers	*(empty)*	*(empty)*	• UT M:P (ACS:VC) • DN	*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*
	Offers to make	*(empty)*	*(empty)*	*(empty)*	• FT P:M (VC:ACS) • UT P:M (VC:ACS) • FT M:P (VC:ACS) • UT M:P (VC:ACS) • DN • FT P:M (VC:DPF) • UT P:M (VC:DPF) • FT M:P (VC:DPF) • UT M:P (VC:DPF)	*(empty)*	*(empty)*	*(empty)*	*(empty)*
	Moral respect O_m	0.75	0.75	0.75 → 0.8125	0.8125	0.8125 → 0.859375	0.859375	0.859375	0.859375
	Moral theory (reasoning)			**Vc →** • **FT P:M (VC:ACS) = 0.454348** • DN = 0.4415	**Vc→** • FT P:M (VC:ACS) = 0.389674 • UT P:M (VC:ACS) = 0.262721 • FT M:P (VC:ACS) = 0.311739 • UT M:P (VC:ACS) = 0.210177 • DN = 0.4415 • **FT P:M (VC:DPF) = 0.454348** • UT P:M (VC:DPF) = 0.335859 • FT M:P (VC:DPF) = 0.363479 • UT M:P (VC:DPF) = 0.268687				
DPF	Active	False	False	False	False	True	True	False	False
	Possessions	• Ph: 2 • m: 70	• Ph: 2 • m: 70	• Ph: 2 • m: 70	• Ph: 2 • m: 70	• Ph: 2→3 • m: 70→40	• Ph: 3 • m: 40	• Ph: 3 • m: 40	• Ph: 3 • m: 40
	Received offers	*(empty)*	*(empty)*	*(empty)*	*(empty)*	• FT M:P at t_3 (DPF:VC) • DN	*(empty)*	*(empty)*	*(empty)*
	Offers to make	*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*	• FT P:M (DPF:VC) • UT P:M (DPF:VC) • FT M:P (DPF:VC) • UT M:P (DPF:VC) • DN • FT P:M (DPF:ACS) • UT P:M (DPF:ACS) • FT M:P (DPF:ACS) • UT M:P (DPF:ACS)		
	Moral respect O_m	0.75	0.75	0.75	0.75	0.75 → 0.8125	0.8125	0.8125	0.8125
	Moral theory (reasoning)					**Vc→** • **FT M:P at t_3 (DPF:VC) = 0.9375** • DN = 0.75	**Vc→** • **FT P:M (DPF:VC) = 0.9375** • UT P:M (DPF:VC) = 0.552978 • FT M:P (DPF:VC) = 0.75 • UT M:P (DPF:VC) = 0.442383 • DN = 0.75		

						• FT P:M (DPF:ACS) = 0.820312 • UT P:M (DPF:ACS) = 0.380859 • FT M:P (DPF:ACS) = 0.65625 • UT M:P (DPF:ACS) = 0.304687		
Moral scenario									
Scheduling		*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*
Execution		*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*
Conclusion		*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*
Entity ACS									
Scheduling		• DN at t_1	• **UT M:P (ACS:VC) at t_2** • UT M:P (ACS:DPF) at t_2	*(empty)*	*(empty)*	*(empty)*	*(empty)*	• DN at t_4	• UT P:M (ACS:VC) at t_5 • **UT P:M (ACS:DPF) at t_6**
Execution		• DN at t_1	*(empty)*	• UT M:P (ACS:VC) at t_2	*(empty)*	*(empty)*	*(empty)*	• DN at t_4	*(empty)*
Conclusion		• DN at t_1	*(empty)*	• UT M:P (ACS:VC) at t_2	*(empty)*	*(empty)*	*(empty)*	• DN at t_4	*(empty)*
Entity VC									
Scheduling		*(empty)*	*(empty)*	• FT P:M (VC:ACS) at t_2	• **FT P:M at t_3 (VC:DPF)**	*(empty)*	*(empty)*	*(empty)*	*(empty)*
Execution		*(empty)*	*(empty)*	• FT P:M (VC:ACS) at t_2	*(empty)*	• FT P:M at t_3 (VC:DPF)	*(empty)*	*(empty)*	*(empty)*
Conclusion		*(empty)*	*(empty)*	• FT P:M (VC:ACS) at t_2	*(empty)*	• FT P:M at t_3 (VC:DPF)	*(empty)*	*(empty)*	*(empty)*
Entity DPF									
Scheduling		*(empty)*	*(empty)*	*(empty)*	*(empty)*	• FT M:P at t_3 (DPF:VC)	• **FT P:M (DPF:VC) at t_5**	*(empty)*	*(empty)*
Execution		*(empty)*	*(empty)*	*(empty)*	*(empty)*	• FT M:P at t_3 (DPF:VC)	*(empty)*	*(empty)*	*(empty)*
Conclusion		*(empty)*	*(empty)*	*(empty)*	*(empty)*	• FT M:P at t_3 (DPF:VC)	*(empty)*	*(empty)*	*(empty)*

After finishing the first four time-frames, let us have a look at what is going on in the next four.

Time		t_5		t_6		t_7		t_8	
Entity	**Observables**	$t_{5'}$	$t_{5''}$	$t_{6'}$	$t_{6''}$	$t_{7'}$	$t_{7''}$	$t_{8'}$	$t_{8''}$
ACS	Active	False	False	False	False	True	True	False	False
	Possessions	• Ph: 2 • m: 100	• Ph: 2 • m: 100	• Ph: 2 • m: 100	• Ph: 2 • m: 100	• Ph: 2 • m: 100	• Ph: 2 • m: 100	• Ph: 2 • m: 100	• Ph: 2 • m: 100
	Received offers	*(empty)*	*(empty)*	*(empty)*	*(empty)*	• DN	*(empty)*	*(empty)*	*(empty)*
	Offers to make	*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*	• FT P:M (ACS:VC) • UT P:M (ACS:VC) • FT M:P (ACS:VC) • UT M:P (ACS:VC) • DN • FT P:M (ACS:DPF) • UT P:M (ACS:DPF) • FT M:P (ACS:DPF)	*(empty)*	*(empty)*

							• UT M:P (ACS:DPF)		
	Moral respect O_m	0.375	0.375	0.375	0.375	0.375	0.375	0.375	0.375
	Moral theory (reasoning)	*(empty)*	*(empty)*	*(empty)*	*(empty)*	Vc→ • DN = 0	**Vc→** • FT P:M (ACS:VC) = 0.263061 • **UT P:M (ACS:VC) = 1.25** • FT M:P (ACS:VC) = 0.210449 • UT M:P (ACS:VC) = 1 • DN = 0 • FT P:M (ACS:DPF) = 0.246582 • **UT P:M (ACS:DPF) = 1.25** • FT M:P (ACS:DPF) = 0.197266 • UT M:P (ACS:DPF) = 1	*(empty)*	*(empty)*
VC	Active	True	True	False	False	False	False	True	True
	Possessions	• Ph: 1→2 • m: 80→50	• Ph: 2 • m: 50	• Ph: 2→1 • m: 50→80	• Ph: 1 • m: 80	• Ph: 1 • m: 80	• Ph: 1 • m: 80	• Ph: 1→2 • m: 80→50	• Ph: 2 • m: 50
	Received offers	• FT M:P (VC:DPF) • DN	*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*	• FT M:P (VC:ACS) • FT M:P (VC:DPF) • DN	*(empty)*
	Offers to make	*(empty)*	• FT P:M (VC:ACS) • UT P:M (VC:ACS) • FT M:P (VC:ACS) • UT M:P (VC:ACS) • DN • FT P:M (VC:DPF) • UT P:M (VC:DPF) • FT M:P (VC:DPF) • UT M:P (VC:DPF)	*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*	• FT P:M (VC:ACS) • UT P:M (VC:ACS) • FT M:P (VC:ACS) • UT M:P (VC:ACS) • DN • FT P:M (VC:DPF) • UT P:M (VC:DPF) • FT M:P (VC:DPF) • UT M:P (VC:DPF)
	Moral respect O_m	0.859375 → 0.89453125	0.89453125	0.89453125 → 0.9208984375	0.9208984375	0.9208984375	0.9208984375	0.9208984375 → 0.940673828125	0.940673828125
	Moral theory (reasoning)	**Vc →** • **FT M:P (VC:DPF) = 0.462884** • DN = 0.404286	**Vc→** • FT P:M (VC:ACS) = 0.389674 • UT P:M (VC:ACS) = 0.262721 • FT M:P (VC:ACS) = 0.311739 • UT M:P (VC:ACS) = 0.210177 • DN = 0.4415 • **FT P:M (VC:DPF) = 0.469285** • UT P:M (VC:DPF) = 0.358201 • FT M:P (VC:DPF) = 0.375428 • UT M:P (VC:DPF) = 0.286561	*(empty)*	*(empty)*	*(empty)*	*(empty)*	Vc→ • FT M:P (VC:ACS) = 0.389674 • **FT M:P (VC:DPF) = 0.474086** • DN = 0.4415	**Vc→** • FT P:M (VC:ACS) = 0.389674 • UT P:M (VC:ACS) = 0.262721 • FT M:P (VC:ACS) = 0.311739 • UT M:P (VC:ACS) = 0.210177 • DN = 0.4415 • **FT P:M (VC:DPF) = 0.477687** • UT P:M (VC:DPF) = 0.370769 • FT M:P (VC:DPF) = 0.382150 • UT M:P (VC:DPF) = 0.296615
DPF	Active	False	False	True	True	False	False	False	False
	Possessions	• Ph: 3→2 • m: 40→70	• Ph: 2 • m: 70	• Ph: 2→3 • m: 70→40	• Ph: 3 • m: 40	• Ph: 3 • m: 40	• Ph: 3 • m: 40	• Ph: 3→2 • m: 40→70	• Ph: 2 • m: 70
	Received offers	*(empty)*	*(empty)*	• FT M:P (DPF:ACS) • FT M:P (DPF:VC) • DN	*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*
	Offers to make	*(empty)*	*(empty)*	*(empty)*	• FT P:M (DPF:VC) • UT P:M (DPF:VC) • FT M:P (DPF:VC) • UT M:P	*(empty)*	*(empty)*	*(empty)*	*(empty)*

					(DPF:VC) • DN • FT P:M (DPF:ACS) • UT P:M (DPF:ACS) • FT M:P (DPF:ACS) • *UT M:P (DPF:ACS)*				
	Moral respect O_m	0.8125 → 0.859375	0.859375	0.859375 → 0.89453125	0.89453125	0.89453125	0.89453125	0.89453125 → 0.9208984375	0.9208984375
	Moral theory (reasoning)	*(empty)*	*(empty)*	**Vc→** • FT M:P (DPF:ACS) = 0.820312 • **FT M:P (DPF:VC) = 0.9375** • DN = 0.75	**Vc→** • **FT P:M (DPF:VC) = 0.9375** • UT P:M (DPF:VC) = 0.567398 • FT M:P (DPF:VC) = 0.75 • UT M:P (DPF:VC) = 0.453918 • DN = 0.75 • FT P:M (DPF:ACS) = 0.820312 • UT P:M (DPF:ACS) = 0.380859 • FT M:P (DPF:ACS) = 0.65625 • UT M:P (DPF:ACS) = 0.304687	*(empty)*	*(empty)*	*(empty)*	*(empty)*
Moral scenario									
Scheduling		*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*
Execution		*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*
Conclusion		*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*	*(empty)*
Entity ACS									
Scheduling		*(empty)*	*(empty)*	*(empty)*	*(empty)*	• DN at t_7	• **UT P:M (ACS:VC) at t_8** • UT P:M (ACS:DPF) at t_9	*(empty)*	*(empty)*
Execution		*(empty)*	*(empty)*	*(empty)*	*(empty)*	• DN at t_7	*(empty)*	*(empty)*	*(empty)*
Conclusion		*(empty)*	*(empty)*	*(empty)*	*(empty)*	• DN at t_7	*(empty)*	*(empty)*	*(empty)*
Entity VC									
Scheduling		• FT M:P (VC:DPF) at t_5	• FT P:M (VC:DPF) at t_6	*(empty)*	*(empty)*	*(empty)*	*(empty)*	• FT M:P (VC:DPF) at t_8	• FT P:M (VC:DPF) at t_9
Execution		• FT M:P (VC:DPF) at t_5	*(empty)*	• FT P:M (VC:DPF) at t_6	*(empty)*	*(empty)*	*(empty)*	• FT M:P (VC:DPF) at t_8	*(empty)*
Conclusion		• FT M:P (VC:DPF) at t_5	*(empty)*	• FT P:M (VC:DPF) at t_6	*(empty)*	*(empty)*	*(empty)*	• FT M:P (VC:DPF) at t_8	*(empty)*
Entity DPF									
Scheduling		*(empty)*	*(empty)*	• FT M:P (DPF:VC) at t_6	• FT P:M (DPF:VC) at t_8	*(empty)*	*(empty)*	*(empty)*	*(empty)*
Execution		*(empty)*	*(empty)*	• FT M:P (DPF:VC) at t_6	*(empty)*	*(empty)*	*(empty)*	• FT P:M (DPF:VC) at t_8	*(empty)*
Conclusion		*(empty)*	*(empty)*	• FT M:P (DPF:VC) at t_6	*(empty)*	*(empty)*	*(empty)*	• FT P:M (DPF:VC) at t_8	*(empty)*

3.3.1.3 Results

At last, after following 8 time-frames we can make a study on the results of the scenario simulation. We can do this by tracking several key observables throughout the time-frames, to see how the different entities fared through time and with their different moral theories.

On the graphs below you can find performance of moral respect (i.e. reputation), number of received offers per time-frame, and the amount of phones and money the entities have.

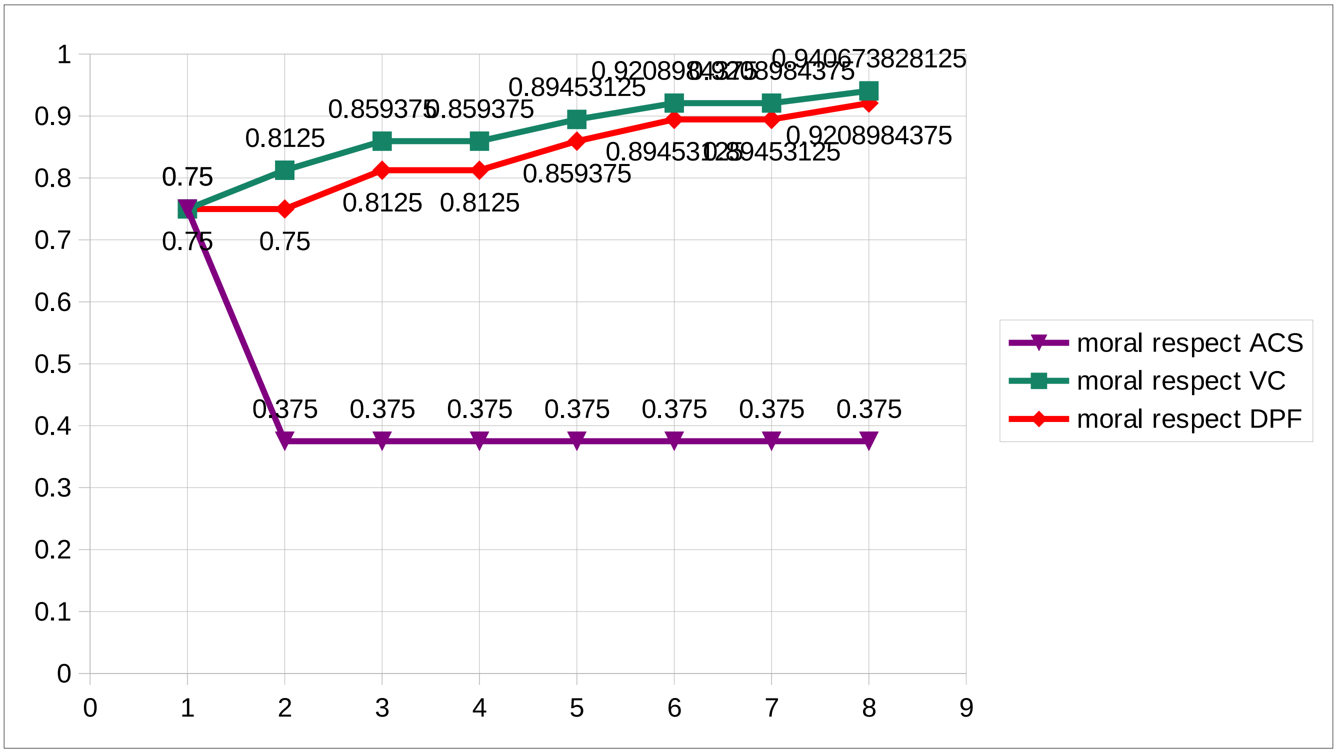

Figure 3: Moral respect for each entity in the Trust and Trade scenario

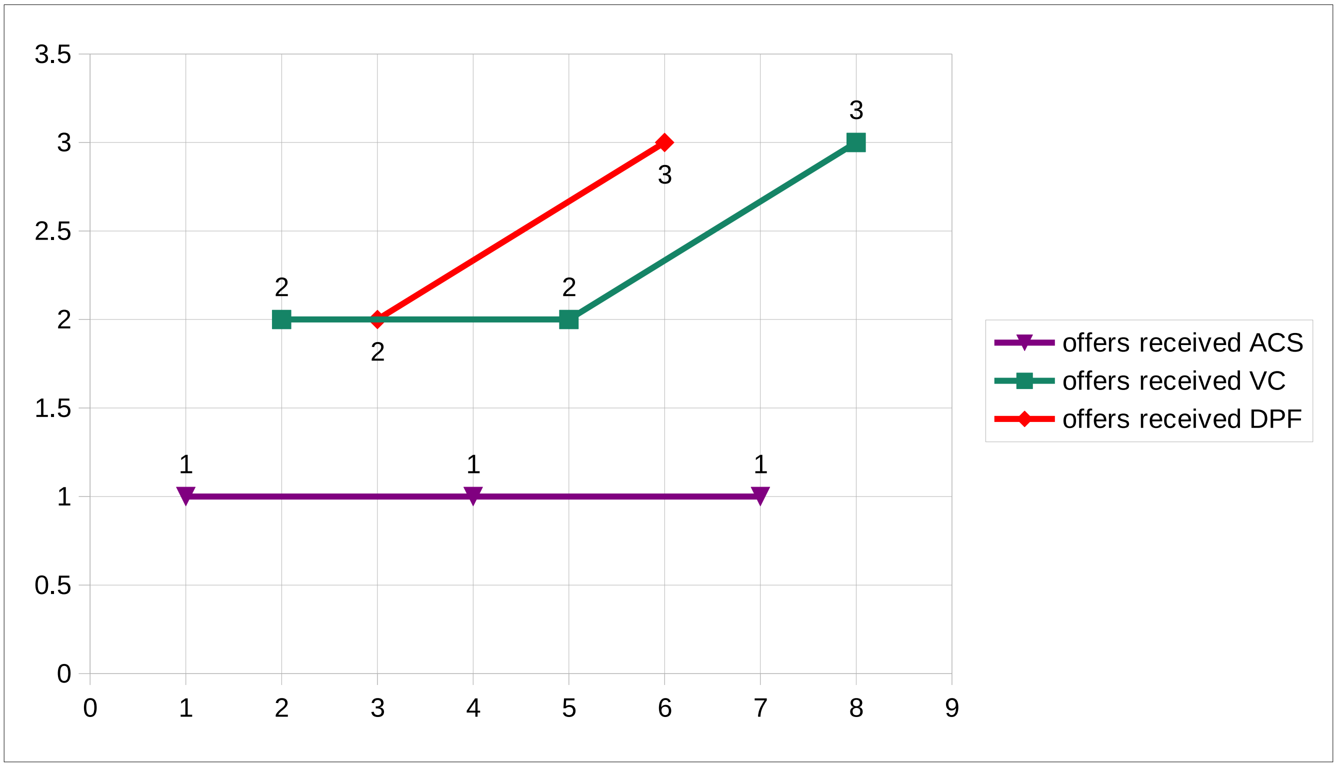

Figure 4: Offers received by each entity per time-frame in the Trust and Trade scenario

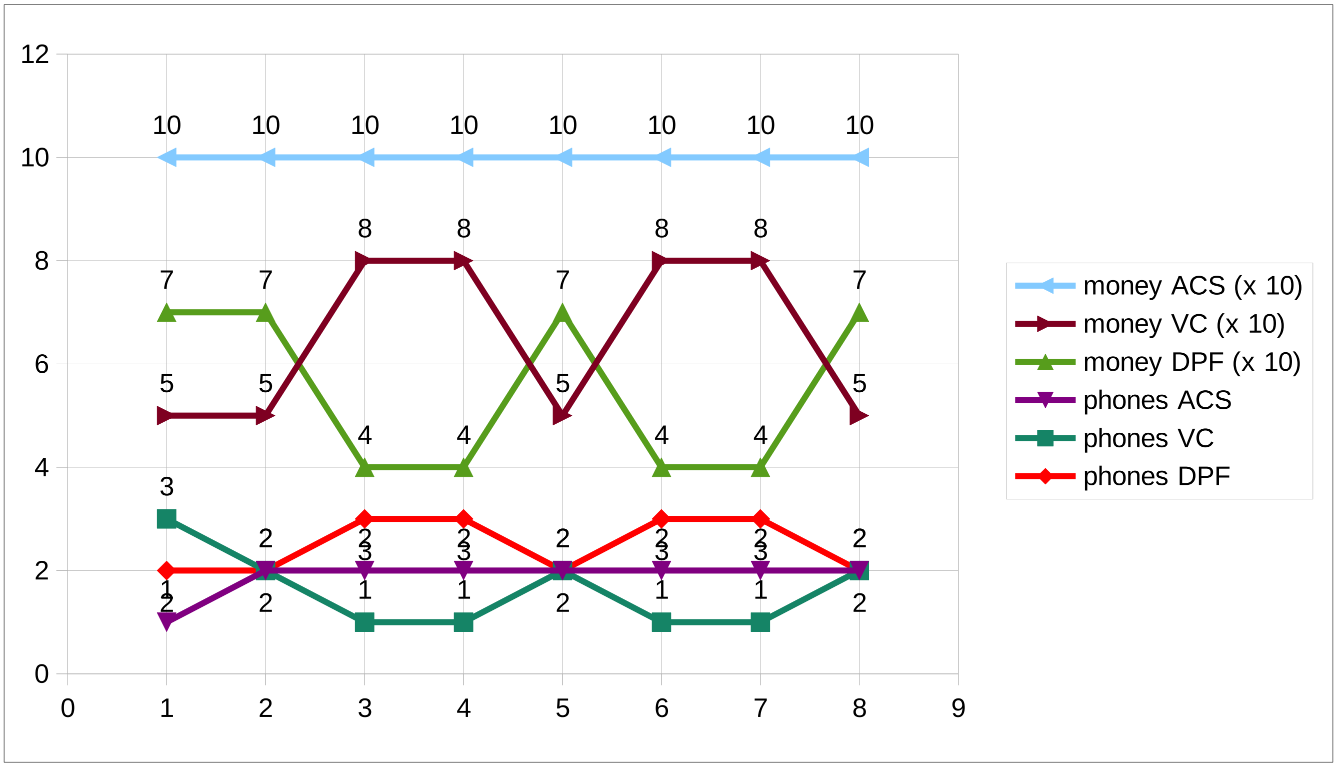

Figure 5: Amount of phones and money for each entity in the Trust and Trade scenario

Tracking the key observables as I did above gives us some key insight about the performance of each entity, as well as, the capacities of the Ethics of Systems Framework and Interface.

We can see that after the initial short-sighted selfish gain made by the selfish entity ACS, by which its reputation (moral respect; O_m) is significantly reduced, this entity is ignored and does not receive any more offers to trade from another entity (except the technical *do nothing* process, simply served by the scenario). Meanwhile, entities VC and DPF continue to trade and steadily increase their reputation because they follow through on what they say they will do. Or, in other words, they walk the walk besides talking the talk. It seems that it pays to be moral and work on building and maintaining one's reputation—or at the very least, to play fair, even if this is in one's selfish interest.

We can also see that VC and DPF continue to exchange money for phones. With this we can make actual economical simulation aided with a moral dimension, since the decisions to trade are made by the entities based on both their internal moral theory and their needs (i.e. for a phone or money) in the particular moment. Since they both find themselves on the counter-side of what they need, they repeatedly enter into trade with one another.

Interestingly enough, ACS would have almost been engaged again in trade with DPF at t_6 (as its offer to DPF was higher-rated than the *do nothing* process). However, DPF also received an offer to trade from the higher-reputation VC, and, wisely enough, decided to take that one instead.

Here we can have a peek at monopolistic practices and why entities are motivated to develop them. Namely, if there was no competition at t_6, the low-reputation ACS would have been picked because DPF needed to trade. However, competition allowed DPF to pick a higher-reputation entity. Therefore, ACS would be all too happy to eliminate its competition so that it doesn't have to work hard to maintain its reputation and can serve products and services with decreasing quality, while with increasing prices. However, this trend would soon surpass the tolerance potential trading partners have, after which they will opt to not trade at all instead of pursuing a bad deal with a monopolist. It is my intuition that similar phenomena develop also in the sphere of the political, as well as any other societal domain.

From the technical side of things, we can see that EoS Framework and Interface are capable of modeling and simulating complex moral scenarios with multiple entities, internal decision-making, reputation tracking, and in general, tracking the whole moral process. This was the end goal of my work here, which I believe have successfully demonstrated.

Making the simulation more realistic

To make the simulation more realistic (and, of course, computationally more demanding), we can increase the number of products and services on offer, diversify needs, and include a higher number of entities that will have slightly different moral theories (e.g. slightly different rule pertinence values and importance of moral rules).

We can additionally try to simulate how communal moral reputation (moral respect) actually works in real life. An example would be where moral respect catches up to an entity gradually and after several time-frames, instead of immediately. Also, moral respect that pertains to a particular entity can be differentially represented in the subjectivity of the other entities, because in order to modify it they would need either to directly attest to the behavior of the subject entity, or receive messages from other entities, which, furthermore, would be judged on their own reputation, etc.

The simulation can be also more realistic if we draw upon findings from representative human cohorts on the importance of moral rules in the different theories, how much particular (classes of) processes pertain to them, and also in varying these on a per-entity basis. This is discussed in the next two chapters.

4 Conclusion

The purpose of this chapter is to demonstrate the capacities and applicability of the Ethics of Systems Framework and Interface to ethics in general, and of course, to AI ethics.

To achieve this, I use two moral scenarios—the classic Trolley problem and the Trust and Trade scenario—as well as 4 classes of ethical theories: consequentialism, deontology, virtue ethics, and EoS' own four ethical principles. All the aforementioned are designed, implemented, and simulated by using the EoS Interface.

The different moral theories offer different answers as to the best course of action in the moral scenarios. This reflects common moral reasoning in humans, which can be represented in a logico-computational manner in order to be implemented in digital systems (e.g. AI entities). These differing answers are delivered after the complex interplay of the theories' internal reasoning mechanisms, which result in assigning particular choice values to each available moral process at the time of consideration. Modifying these internal cause-effect mechanisms, as well as rule pertinence values of moral processes, would bring different results which can at times radically change the answers. Therefore, care ought be taken to formulate these internal mechanisms, external interpretations and representations, as well as, the holistic representation of the moral scenarios themselves, in order to ensure morally-sound decision-making on the part of AI systems.

With the demonstration of the capacities of the EoS Framework and Interface, I assert that the purpose of this chapter is achieved.

Ethics of Systems is a novel approach that can make significant contribution towards understanding, modeling and managing moral scenarios in general. Of course, this is not to say that each and every aspect of ethics and morality can be represented in a conceptual, computational, or even explicit manner at all.

However, given that AI entities increasingly take upon roles in our societies that have increasing moral relevance—all the while steadily escaping our direct command and control—we will need to provide them with the means to understand our moral perspectives and make sure to respect them, or even perform better than we would ourselves.

Ethics of Systems Framework and its Interface, alongside other ones, can be taken as the right tools for this endeavor.

4.1 Implications carried forward

There are some implications arising from this chapter and Chapter III. Towards Ethics of Systems (the Metaethics) that I will discuss in the next Chapter V. Discussion.

Chapter V. Discussion

1 Introduction

In this chapter I am discussing whether the two main research questions, as well as the sub-questions, presented in the first chapter (see 3 Research objectives and questions in Chapter I. Introduction) are answered in this thesis. The substantial research effort in this thesis is focused on the creation of a novel theory (framework) capable of modeling and managing moral scenarios in which AI entities participate. If you remember, a candidate theory that provides a satisfactory answer to the research questions would be able to formally represent moral scenarios (and their components and attributes) in which AI entities participate, provide the means to perform moral calculus on available courses of (in)action, and in accordance to formally-designed moral theories.

For this purpose, it helps to restate the questions and sub-questions here. The two main research questions are:

- What theory can explain moral scenarios in which AI entities are participants?
- What theory can explain the process of moral reasoning, decision and action for AI entities in virtual, simulated and real-life moral scenarios?

The research sub-questions are:

- What are the major ethical issues raised by the introduction of AI entities in the world and in human societies?
- What are the foundational systemic and informational attributes of moral scenarios whose participants include AI entities?
- Which are the foundational ethical principles, concepts and methods of reasoning relevant to AI entities?
- Can the foundational ethical principles, concepts and methods of reasoning relevant to AI entities be systematized into a coherent (and possibly comprehensive) ethical framework?
- In what way can such ethical framework provide means for, or assist, reasoning in moral scenarios in a morally-sound manner?
- Can such an ethical framework be translated or paraphrased into legal, technical, engineering, and other instruments?
- What are the ethical, scientific, and possibly legal implications that this kind of a comprehensive study brings on AI ethics, and ethics in general?

The aforementioned are discussed in the sections that follow.

2 Discussion

The research work presented here is undertaken in two distinct halves of the whole research effort which build upon one another.

The first half is focused on the derivation of a novel (meta)ethical framework for AI ethics, and for ethics in general. This (meta)ethical framework is named **Ethics of Systems**. Alongside the framework itself, this first half has also resulted in the derivation of its main methodological tool: the **Ethics of Systems Interface**.

The second half is focused on testing this (meta)ethical framework by applying it to two hypothetical moral scenarios in which AI entities are participating in active roles (i.e. decision-making and executing) and passive roles (i.e. as recipients of moral effects). This part of the research effort has the purpose of testing—and through this, demonstrating—the capacities of **Ethics of Systems** Framework, as well as of its methodological tools (e.g. the Ethics of Systems Interface, the Method of Levels of Abstraction, methods from systems science, and others), when applied to moral scenarios, in particular, those that include AI alongside non-AI entities as their participants.

These two distinct halves of the research effort result in the two chapters of this thesis, Chapter III. Towards Ethics of Systems (the Metaethics) and Chapter IV. Applying Ethics of Systems Framework to AI moral scenarios (the Ethics). Their developments and results are commented below.

2.1 Towards Ethics of Systems

The first half of the research effort, presented in Chapter III. Towards Ethics of Systems (the Metaethics), is directed towards establishing the foundations of a novel (meta)ethical framework for AI entities, named the **Ethics of Systems**.

Ethics of Systems draws upon research and findings from several relevant major fields: ethics, ethics of AI, philosophy and ethics of information, and systems theory. It also draws upon, or aims to be compatible with, research from many other relevant fields: computer science, object-oriented programming, law, argumentation theory, decision theory, panpsychism, theory of integrated information, logic, and basic set theory.

Then, by using the methodological tools present in philosophy and ethics of information, and systems theory, I formulate a formal framework which can represent moral scenarios where AI entities are included in a consistent and coherent manner.

This results in the discovery of the main systemic and moral striving of systems (entities): the maximal increase of their **Quality of Life**—which in other words can be defined as the state of matters of their **flourishing**. Quality of Life is, furthermore, discovered to be the product of two axiomatic major goals, known as moral imperatives: the **Achievement of Personal Goals** (APG), and the **Conservation of Personal Continuum** (CPC).

In Ethics of Systems' axiology, the state of flourishing is defined as the moral **Good**; and the major obstacle to flourishing—destructive metaphysical entropy—is defined as its moral opposite: the moral **Bad**. If destructive metaphysical entropy is introduced intentionally, this belongs to the subset of moral Bad defined as moral **Evil**. On the other hand, if destructive metaphysical entropy is not introduced intentionally, this belongs to the subset defined as moral **tragedy**.

With this in mind, I define the four basic ethical principles of Ethics of Systems, whose aim is to guide morally-relevant behavior of all systems towards the Good of systems and the systemsphere in general, and strive away from moral bad. These four principles are:

0 Destructive entropy ought not to be caused in the systemsphere (null law)

1 Destructive entropy ought to be prevented in the systemsphere

2 Destructive entropy ought to be removed from the systemsphere

3 The flourishing of systems as well as of the whole systemsphere ought to be promoted by preserving, cultivating, and enriching their well-being

These four principles are reformulation of Floridi’s own four basic ethical principles of ethics of information (Floridi, 2013; p. 71), adapted to the particularities of Ethics of Systems.

This first half of the whole research effort also result in the development of the foundational methodological tool of Ethics of Systems, namely: its **Interface**. The EoS Interface is a formalized approach at modeling **moral scenarios**. It can represent their components (i.e. **moral and other entities**, **moral theories**), relations (i.e. **moral and other processes**, hierarchies and ontologies), as well as other relevant but non-systemic phenomena, such as time, space and location, and anything else of relevance to a particular scenario.

The purpose is to devise a formal method of representing the philosophical substance behind the Ethics of Systems Framework, which formal representation can then be applied directly to moral scenarios. The ultimate goal is, of course, to enable explicit modeling and managing of moral scenarios where AI entities, and by extension all other entities, are participants.

The work in Chapter III. Towards Ethics of Systems (the Metaethics) results in the creation of both the Ethics of Systems Framework and its Interface. With this in mind, Chapter III. Towards Ethics of Systems (the Metaethics) provided a direct answer to the following research sub-questions:

- What are the foundational systemic and informational attributes of moral scenarios whose participants include AI entities?

  Answer: these are *moral entities*, *moral processes*, *moral theories*, *moral Good* (flourishing of systems and the systemsphere), *moral Bad* (destructive metaphysical entropy), *Quality of Life* as a measure of flourishing, the two moral imperatives: *Achievement of Personal Goals* and *Conservation of Personal Continuum*, and *agency* and *patiency*.

- Which are the foundational ethical principles, concepts and methods of reasoning relevant to AI entities?

  Answer: The most foundational ones of these, discovered and established by this thesis, are *Quality of Life*, and the two moral imperatives: *Conservation of Personal Continuum* and *Achievement of Personal Goals*.

- Can the foundational ethical principles, concepts and methods of reasoning relevant to AI entities be systematized into a coherent (and possibly comprehensive) ethical framework?

  Answer: yes. One such effort is the Ethics of Systems framework.

- In what way can such an ethical framework provide means for, or assist, reasoning in moral scenarios in a morally-sound manner?

  Answer: by explicitly representing moral phenomena in a moral scenario, and by providing the formal means to design, apply and follow moral theories, Ethics of Systems Framework and Interface enable performing of moral calculus by moral entities on available moral processes ((in)actions) in moral scenarios.

- Can such an ethical framework be translated or paraphrased into legal, technical, engineering, and other instruments?

  Answer: yes. The Ethics of Systems Interface is a formalized approach at representing the Framework, and establishes the possibility to be paraphrased into other formalized approaches

> (e.g. logic, mathematics, legal and procedural instruments, formal ethical approaches, decision and argumentation theory, programming and digital systems design and engineering, etc.). Additionally, this process can work in the inverse way i.e. legal, ethical and engineering formalizations can be paraphrased into the language of the Interface, and through that, in the language of the Framework itself.

The second part of the research effort, however, is explored through in the following Chapter IV. Applying Ethics of Systems Framework to AI moral scenarios (the Ethics).

2.2 Ethics of Systems and AI ethics

The work in the previous Chapter III. Towards Ethics of Systems (the Metaethics) provides answers to many of the research sub-questions. However, the main research questions are not answered there. The reason for this is that even though the Ethics of Systems Framework and its Interface are designed in that chapter, they are not demonstrated as capable of performing as stated, and therefore do not yet demonstrate my claims.

This is the work of the following chapter, Chapter IV. Applying Ethics of Systems Framework to AI moral scenarios (the Ethics). The purpose of this chapter is to demonstrate the capacity of the Ethics of Systems Framework and its Interface in modeling moral scenarios and all their components, especially where AI entities are participants.

For this purpose, I use the EoS Interface to first translate and design four classes of ethical theories: **consequentialism**, **deontology**, **virtue ethics**, and **EoS' own four ethical principles**. In total, there are **8 different theories modeled**. After this, the Interface is used again to model two moral scenarios: the **classic Trolley problem** and the **Trust and Trade scenario** (a turn-based trading simulation). The 8 different theories are embedded in different moral entities in the two scenarios and used to guide their decision-making. As expected, the behavior of participants is dynamic and different. It is the consequence of it being based on circumstance and their embedded moral theories, but coherent with expectations how those theories should perform.

The EoS Interface comfortably accommodates all of the above. It does this by providing the necessary formal approach on one hand, and contextual flexibility on another, to successfully model moral scenarios and their components, and track their development—all in an explicit fashion.

Taking this in consideration, as well as the contribution of the work in the previous chapter (see above), the whole thesis work provides the following answers to the two main research questions:

- What theory can explain moral scenarios in which AI entities are participants?

  > **Answer**: such a theory is Ethics of Systems[66] with its main methodological tool, the Ethics of Systems Interface.

- What theory can explain the process of moral reasoning, decision and action for AI entities in virtual, simulated and real-life moral scenarios?

  > **Answer**: such a theory is Ethics of Systems, in combination with approaches that model moral and general decision-making (e.g. decision theory, argumentation theory, axiology, etc.).

This thesis, however, provides only a partial answer to these sub-questions:

[66] This is not to say that Ethics of Systems is the *only* such theory—it is one of many. However, a significant effort was provided in order to make EoS as versatile and applicable to AI ethics as it possibly can, while being accommodating or cooperative with other approaches in AI ethics.

- What are the major ethical issues raised by the introduction of AI entities in the world and in human societies?

  Partial answer: the ethico-philosophical exploration in Chapter III. Towards Ethics of Systems (the Metaethics) and the literature review in Chapter II. Literature review cover many significant issues in regards of the widespread introduction of AI entities. However, this is not the main focus of the study, and hence this is not done in a systematic or in-depth way. There are better studies in this respect, some of which are mentioned in 5 Ethics of AI section in Chapter II. Literature review.

- What are the ethical, scientific, and possibly legal implications that this kind of a comprehensive study brings on AI ethics, and ethics in general?

  Partial answer: there are three kinds of implications: ethical (substantial), technical, and scientific. They are discussed in the following section (see below).

These sub-questions will need to be comprehensively answered in future work, and in research performed by other researchers as well.

2.3 Implications

In this section I will discuss some of the implications that my study uncovers, in particular, those which tended to arise repeatedly. I will start with the substantial (i.e. ethico-philosophical) implications, followed by the technical, and finishing with the scientific ones.

2.3.1 Substantial (ethical) implications

During the research several ethico-philosophical implications were repeatedly arising. These are: complexity, heuristics and pervasiveness of bias; epistemological limitations and derived ontological issues (e.g. implicit moral phenomena); the usefulness of the moral veil of ignorance; and the validity of the systemic level of abstraction when studying moral phenomena. Let’s get to them in order.

Complexity, heuristics, and pervasiveness of bias

When discussing complexity, the usual issues under discussion are the first-hand, primary ones. Examples of these are the inability to pass above a particular upper bound of computational, cognitive or representational power, which necessarily weakens any knowledge method including scientific methods. This, for overly complex situations, in turn, converts coherent and sound knowledge-acquiring methods to their weaker, but more pragmatic, counterpart—heuristics.

However, this epistemological issue is rarely discussed as the birth point of another, very importance ethical phenomenon: bias. Biological entities have inherent cognitive limitations which force them to rely on heuristics. Since humans are also biological entities we are similarly limited in our cognitive capacity, and this translates into limitation when considering moral situations as well. This is the very reason why biases (of which heuristics are one type) are pervasive, since they offer a quick and pragmatic, if not sound, method of obtaining information about a particular situation and making more often than not *good enough* decisions. Similarly, this is the reason why there are ‘weakened’ modes of reasoning such as abduction, probabilistic reasoning, defeasible inferences, and informal logic.

Using digital systems and tools (e.g. AI entities, simulations etc.) can help us significantly reduce bias. However, it seems doubtful that bias will ever be totally eradicated. What digital systems can contribute, instead, is to increase our upper bound of cognitive (computational) capacity when considering moral scenarios. This has an upper limit, however: the Bremermann's computational limit (see 3.4.2 Complexity in Chapter III. Towards Ethics of Systems (the Metaethics)). This limit for now seems insurmountable.

Epistemological limitations and derived ontological issues

This ties in with the next issue of epistemological limitations and resulting ontological problems, such as implicit moral phenomena. The study performed here does, indeed, result in a formalized and explicit approach at AI ethics. However, we should be wary and not assume that *each and every aspect* of ethics can be represented in such a way. The reason is that many of the relevant ethical phenomena don't support formalization or even explicit expression.

Examples would be the so-called esoteric ethics (Copp, 2006; p. 640), error theory (Hussain, 2006), Moore's Open Question Argument (Baldwin, 2010), the ineffable experience of deep meditation as attested by Buddhist practitioners, or of deep prayer and connection with God by esoteric Christianity, or the inexpressible wholeness of the Dao in Daoism. In each of these cases there are (so their proponents claim) some kinds of moral implications arising from them, but they are difficult—or outright impossible—to be formalized and explicitly expressed, manipulated, and explored.

Furthermore, from an internal, subjective (psychological) perspective, humans rarely seem in complete awareness of their internal cognitive mechanisms that guide them to make certain decisions. Many of these have been shaped and hardwired by evolutionary pressures, and are subject of study by fields such as developmental, moral and evolutionary psychology. This is the first outright difficulty when attempting to model and emulate *human* moral decision-making. The second is more indirect, and can appear in various digital systems designed and used by humans. Namely, AI systems are currently predominantly designed by human programmers and system designers, and they can and often will transfer their biases and cognitive shortcomings on these systems even without being aware of it.

And the third issue is with the effect epistemological limitations have on ontological commitments. Namely, if we cannot even begin to explicitly express or formalize a particular moral phenomenon, we are unable to describe it as would be required by an ontological framework, for example. This would mean that the power of our ontological framework will be necessarily weakened, and again, turned into a semi-heuristic method by introducing defeasibility of the inferences that can be made by using it. Or, in other words, if AI systems cannot fully understand us or other systems, they will run the risk of making error-prone, morally-contentious and at times even morally-abhorrent decisions that have moral effects on us and other systems.

The usefulness of the moral veil of ignorance

The moral veil of ignorance was recently reintroduced and made famous by John Rawls in his book, *A Theory of Justice* (Rawls, 2002, 1997, 1971), although the idea was introduced since the eighteen century by proponents such as Kant, Hobbes, Locke, Russeau, and others. However, there are additional arguments in favor of using the moral veil of ignorance besides those offered by Rawls himself.

Namely, this approach states that we ought ignore (irrelevant) details about the participants in a particular scenario, of which we could be the one receiving the effects, and make a decision based on that. This informationally 'stripped-down' approach can help reduce bias, since if it is paired with randomized decision-making for choices with equal choice value (i.e. as the one I used in 2.1.1 Algorithmic decision flowchart in Chapter IV. Applying Ethics of Systems Framework to AI moral scenarios (the Ethics)) it will tend to give equal chance to every involved entity in a scenario from the cohort of moral processes to be picked at any particular time. Additionally, it would help reduce complexity of moral situations, which in turn can increase reaction times, as well as, reduce epistemological issues (e.g. data-gathering and data bias).

An example would be where an autonomous vehicle would not consider whether the person on the left, right or inside is an old or young person, male or female, animal or human, etc. Instead, it simply would make a randomized choice, or a reduced randomized choice (where the scope of stochasticity is confined within the bounds of the best possible ‘in-between’ plane of choice palette).

Validity of the systemic level of abstraction for studying moral phenomena

Moral scenarios, entities, processes, theories, and other moral phenomena have certain particularities that set them apart from other worldly phenomena. There are many approaches that aim to explicitly, implicitly, or formally represent these. One such, Floridi’s IE (Floridi, 2013), has been discussed extensively in this thesis.

However, in search of more comprehensive representational method for moral phenomena I have turned to systems science and its way of looking at the world and systems. This proved advantageous, since by using the systemic level of abstraction I was able to express moral phenomena in a formal and explicit fashion, while taking care of all their particularities that are relevant in general or in the particular moral scenario.

Although there are many approaches that attempt the same in the field of formal ethics, based on the research performed here the systemic level of abstraction could prove itself as deserving to be located among the best such ones.

2.3.2 Technical implications

Besides ethical, there are several technical implications that repeatedly appeared while performing the research. These are the following: the jump between form and substance; the flexibility of EoS Interface; and the advantages of digital assistance in scenario simulations.

The jump between form and substance

While designing the moral scenarios (i.e. moral theories, entities, processes, etc.) it is clear that many of the particularities in the formalization are based on interpretation. This provides an easy vector of attack based on arbitrariness i.e. one can easily and fairly criticize why, for example, the importance of a particular rule in a moral theory was set at exactly 0.75 instead of 0.63 or any other number; or why that particular process pertains to that particular rule with strength of 0.7 instead of 0.25, -0.5, etc.

Additional critique based on arbitrariness would necessarily involve the particular design of the moral theories. Their design can be criticized on the grounds of interpretation. An example would be what exactly means for an act consequentialist theory to order available moral processes according to the expected utility to be gained.

This is a fair critique on the *particular* design choices that I used in this work. We have to keep in mind, however, that the goal of Chapter IV. Applying Ethics of Systems Framework to AI moral scenarios (the Ethics) is primarily to *demonstrate* that formal theory design can be performed within the EoS Interface, and not do it in a particular manner supported by formal ethics (which is not unified anyway). Involving this approach would have meant to increase the scope of the thesis well beyond what is required, and would have unnecessarily complicated the research effort.

Again, the technical goal of demonstrating that formal theory design can be done is achieved. This means that the EoS Framework and Interface support the jump from substance to form and vice-versa. Certainly, the particularities of the moral theories can and must be improved with further work, and in collaboration between all concerned stakeholders: researchers in multiple relevant scientific fields (e.g. computer science, ethics, philosophy, law), governments, organizations, corporations, and the public.

Flexibility of the EoS Interface

This research effort is, undoubtedly, complex and in-depth, spanning multiple layers in many research disciplines. Such an approach necessitates flexibility of effort and results. One of the (implicit) goals for the EoS Framework and Interface was exactly to accommodate such approaches in research and application. This is, arguably, achieved in Chapter III. Towards Ethics of Systems (the Metaethics), and is additionally demonstrated by exploring the scenarios in Chapter IV. Applying Ethics of Systems Framework to AI moral scenarios (the Ethics).

The Interface is flexible to accommodate different moral theories, modes of reasoning, events, changes and developments, processes, behavior of systems, and many other non-systemic or amoral phenomena. It can, in theory, be accommodating for approaches such as decision theory, argumentation theory, and any other disciplines that deal with agent-based simulations and decision making e.g. BDI, crowd behavior, etc. This will need to be demonstrated in future work, however.

For now, since it uses simple set theory, logic and arithmetic, and allows for the utilization of any kind of logic (formal and informal approaches) or mathematics (e.g. pre-calculus and calculus), it can theoretically accommodate virtually any kind of reasoning approach. Whether any such approach would be computationally feasible depends on context and requirements.

The advantages of digital assistance in scenario simulations

By exploring (‘running’) the moral scenarios in Chapter IV. Applying Ethics of Systems Framework to AI moral scenarios (the Ethics) by hand it becomes clear that it is not an easy task. It also seems a foolish endeavor, since we already have well-developed tools for agent-based simulation.

Therefore, the best approach here would be to use such tools to implement the design prescribed with the EoS Interface, and then let the scenario run inside a computer. This would be additionally supportive of the claim that EoS Framework and Interface can accommodate AI ethics, since, by definition, the entities inside the computer-assisted simulation would be—AI entities.

Unfortunately, I did not have enough time to implement EoS in simulation tools, something which will remain for future work.

2.3.3 Scientific implications

The results of the research work performed here are aimed not just at aiding practical exploration of AI ethics, but also at aiding scientific research in this field. For example, the EoS Interface is a developed methodological tool which can guide further research in ethics and AI ethics, as well as law, computer science, and all fields where it can be discussed and implemented.

By extension, one research implication arising from this is that moral scenarios *can* be explored in a formal manner of this kind (with the caveat regarding implicit moral phenomena discussed above). As mentioned before, it is my intent that the research here contributes significantly towards the scientific, as well as the practical, fields of AI ethics and ethics in general, and also results in the formation of a new field of scientific inquiry: ethics of systems.

2.4 Limitations

As with any research effort, this thesis has several limitations that I will discuss here.

First, the work here is not, and was never aimed to be, the ‘Holy Grail in AI ethics’, by attempting to answer each and every question that arises in the field, and take into consideration each and every perspective. This is not even possible for a well-developed scientific field with its cohort of researchers, much the less of a single doctoral thesis work, and even less for such a dynamic and upcoming field such as AI ethics. However, there is much to be done to improve the work here, and thus turn EoS into a really useful and applicable tool in the field.

Secondly, EoS is incapable of answering which is the ‘right’ or ‘best’ moral theory generally or in particular context. However, it provides the possibility to benchmark the performance of moral entities with different or equal starting points and with different embedded moral theories. This could provide useful when testing for the best-performing theories in particular contexts, bench-marked according to legal and moral requirements of all involved stakeholders.

And thirdly, this research cannot be said to *precisely* reflect human moral reasoning. The reason for this is that many of the parameters that I specified were given intuitively, and hence arbitrarily, in order to demonstrate the capacities of the EoS Framework and Interface. This is a known weakness of all constructive approaches at (AI) ethics i.e. that they need to be field-tested. However, again, the purpose of the simulations are not to perfectly reflect human moral reasoning, but to demonstrate that such reasoning can be represented using EoS Framework and Interface. The exact values of parameters (which at times change in context) ought be determined with extensive research in the fields of moral psychology and empirical ethics.

3 Impact

AI ethics as a field of research is receiving a significant amount of attention. There are certainly some strong efforts, such as the AIHLEG set up by the European Commission (AIHLEG, 2019), the euRobotics topics group (euRobotics topics group on ethical, legal and socio-economic issues (ELS), 2017), the European Group on Ethics in Science and New Technologies (European Commission, 2018), the AI Now group (Crawford et al., 2016), as well as researchers from the Oxford Internet Institute, the Alan Turing Institute and the Digital Catapult (Morley, Floridi, Kinsey & Elhalal, 2019), the CLAIRE research network (CLAIRE, 2020), and others.

However, we are currently far from reaching a consensus on the basic building blocks in the field i.e. on ethico-philosophical foundations, scope, or methodology. The contribution of this thesis is exactly in this direction. It makes a significant contribution on a methodological level by delivering the EoS Interface. The EoS Interface is a methodological tool that can be used to explicitly and formally represent moral scenarios in a consistent and coherent manner, translatable or paraphrasable across disciplines, authors, and organizations. This thesis is also making a significant contribution on a substantial (ethical) level with the EoS Framework itself, with insight gathered from ethics, AI ethics, systems theory, philosophy and ethics of information. This is aided by the capacity of the EoS Framework and Interface to bridge the gap between form and substance.

4 Future work

In many occasions throughout the text I mentioned that some particularities have to be left for future work. Now is the time to recapitulate them so that future research effort can be targeted to improve the work done here, and of course, the results achieved in this thesis.

The first and most obvious improvement that can be performed on the work here is going beyond the foundational nature of the research. Namely, this thesis was aimed at establishing the *foundations* of an ethical framework for AI entities. The foundations are now set, but they can and should be improved with future effort. This would entail improvement on moral theory design, increase of the complexity of moral scenarios, exploring individual and collective behavior, exploring more case-studies, improving rule pertinence values of moral processes to particular rules in particular contexts, and more.

A second improvement would be work on improving representation of relations between rules. For now, the EoS Framework and Interface simply define a measure of rule importance. However, a fully developed ontology of rules with their prioritization relations—e.g. lex generalis - lex specialis; or argumentation-style attacks (undercutting, rebutting, and undermining) and support—would be a welcome development. This ontology can be 'borrowed' from an outside database and interpreted in a JIT (just-in-time) fashion; or it can be implemented directly by using the EoS Framework.

A third improvement would be in widening the scope of theories designed and tested within the EoS Framework and Interface. This effort focused on four classes of ethical theories—consequentialism, deontology, virtue ethics, and EoS's own four basic ethical principles. It focused on 8 distinct moral theories—act consequentialism, scalar consequentialism, Ross and Audi's *prima facie* duties, Bible's Decalogue, Rawlsian Maximin, classic (agent-focused) virtue ethics, ethics of care (patient-focused virtue ethics), and EoS four ethical principles.

As we can see, we are missing several other distinct ethical theories, such as variants from indirect consequentialism (e.g. rule, motive, and sophisticated consequentialism), environmental ethics (whether this one can be subsumed under ethics of care is a subject of interpretation) and deep ecology, Kantian ethics, rights-based approaches, eastern approaches (e.g. Buddhist, Hinduist, Daoist, Bushido, and Islamic ethics), African approaches (e.g. Ubuntu), collectivist ethics, Eastern European approaches, classical, liberal, conservative and progressive as well as leftist and rightist morality, and other approaches. Theoretically, the EoS Framework and its Interface are capable of representing the aforementioned (with exception for their inexpressible, implicit aspects of which I already discussed before). In order to improve the widespread applicability of EoS, exploring these other approaches can be a fair focus of future work.

A fourth improvement would be the normalization of choice value Vc for all theories within a unified interval. This would, in turn, enable a unified bench-marking effort of the different theories in particular context in order to determine the best-performing ones.

A fifth improvement is to dive deeper into particularities of moral theories and input them in the moral theory design. One example would be allowing for different rule pertinence values, or even different rules, that deal with discrepancy between *doing* (active participation) and *allowing* (passive participation and negligence), which people predictably consider as having different contribution to the (im)morality of a particular moral process.

A sixth improvement would be to successfully design and embed approaches from disciplines such as argumentation theory, decision theory, and agent-based simulations. These can be either performed in an JIT fashion (by importing a particular decision-making module in a moral entity), or can be designed directly by using the EoS Interface.

A seventh improvement, which is a significant one, would be to use simulation software to explore moral scenarios in a significantly improved and manageable fashion. This is an area which I expect to contribute to in the near future myself.

Chapter VI. Conclusion

The purpose of this research endeavor is to create a new theoretical ethical theory (framework) which can provide the means of modeling and managing moral scenarios in which AI entities participate. This is achieved in a two-fold manner.

First, I develop the foundations of a novel ethical framework for AI entities by drawing upon findings from the fields of ethics, ethics of AI, law, philosophy of information, ethics of information, and systems science. This framework I name **Ethics of Systems**. Alongside the Framework itself, I develop its main methodological tool: the **Ethics of Systems Interface**. This Interface provides the formal means of representing the Framework, as well as using it to model, track, and manage moral scenarios where AI and other entities are included as active and passive participants. This first half of the research endeavor is explored in Chapter III. Towards Ethics of Systems (the Metaethics).

Second, the focus of the subsequent Chapter IV. Applying Ethics of Systems Framework to AI moral scenarios (the Ethics) is testing and demonstrating the capacities of the Ethics of Systems Framework. In this chapter, I use the Ethics of Systems Framework and Interface to explore **four classes of moral theories** (consequentialism, deontology, virtue ethics, and Ethics of Systems), **8 moral theories in total** (direct and scalar consequentialism, Ross and Audi *prima facie* duties, Biblical Decalogue and Rawls' Maximin; classic virtue ethics and ethics of care; and Ethics of Systems' own Four ethical principles), and two moral scenarios—the **classic Trolley problem** and the **Trust and Trade scenario**. The Framework and its Interface prove capable to model, track and manage moral scenarios, as well as, provide the means for their dynamic development without the need for direct command and control from an external controller. This concludes the substantive part of this research endeavor.

AI ethics is a field that is about to expand immensely and demand a great deal of our attention. The reason for this is the simple fact that, with their widespread introduction in our societies, the scope of tasks and roles AI systems undertake is about to widen significantly, replacing humans in this process. They will initially replace us in boring, simple and repetitive tasks and roles that no one wants (to pay to be performed), but will soon start replacing us in increasingly more complex and significant ones. Eventually—taking into consideration the exponential nature of technological development—given enough time, there are probably no such roles that can indefinitely resist to this take-over.

Many of these tasks and roles belong, or will belong, to the sphere of the moral. Although societies, states and super-governmental organizations are taking the typical approach of introducing legal regulations that aim to deal with the status and behavior of AI systems and their designers, producers, users, and receivers of effects, this is not sufficient. The reason is that—besides the fact that law is always lagging behind technological advancements—there are many phenomena in this sphere which belong to the so-called soft law and ethics, and which cannot and ought not be outright regulated with hard law.

This is why multidisciplinary research, such as the one performed in this thesis, is crucial. By providing useful insight into the core of AI ethics and thus relevant important issues, while in parallel providing the needed methodological tools for practical application and aiding further scientific research, it helps us, people, and helps the AI entities we employ, to make sense of it all and manage it in a morally-sound manner. Therefore, I expect this work to represent such contribution to the field of AI ethics. I also expect for it to help tip the scales away from the dystopian futures that we so often imagine in our great works of fiction.

And with this being said, we can finally conclude this undertaking.

I sincerely thank you for your time and patience.

Chapter VII. Bibliography

1 References

- euRobotics topics group on ethical, legal and socio-economic issues (ELS) (2017). Ethical, Legal and Socio-economic Issues in Robotics [position paper]. Philosphy & Theory of Artificial Intelligence.
- AIHLEG (2019). Ethics Guidelines for Trustworthy AI. High-Level Expert Group on Artificial Intelligence.
- Alexander, L. & Moore, M. (2016). Deontological Ethics. In E. N. Zalta (Ed.), Deontological Ethics (Winter 2016 ed.). : Metaphysics Research Lab, Stanford University
- Al-Jazzeera (2019). Boeing admits flaws in 737 MAX simulator software after crashes. Retrieved from https://www.aljazeera.com/news/2019/05/boeing-admits-flaws-737-max-simulator-software-crashes-190519054206962.html
- Ambrose, M. L. (2014). The law and the loop. , The law and the loop. : IEEE Press
- Ananny, M. & Crawford, K. (2016). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. new media & society, , 1461444816676645.
- Anderson, M. & Anderson, S. L. (2011). Machine ethics. : Cambridge University Press. Retrieved from https://www.ebook.de/de/product/14736392/machine_ethics.html
- Andrighetto Giulia, Governatori Guido, Noriega Pablo & van der Torre, L. (2013). Normative multi-agent systems (Vol. 4). : Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.
- Annas, J. (2006). In D. Copp (Ed.), Virtue ethics (pp. 515-536). : Oxford University Press Oxford, England.
- Arnold, T. & Scheutz, M. (2016). Against the moral Turing test: accountable design and the moral reasoning of autonomous systems. Ethics and Information Technology, 18(2), 103-115.
- Ashby, W. R. (1999). An Introduction to Cybernetics. : Chapman & Hall Ltd..
- Ashby, W. R. (2001). General systems theory as a new discipline. , General systems theory as a new discipline. : Springer
- Ashrafian, H. (2015b). AIonAI: A humanitarian law of artificial intelligence and robotics. Science and engineering ethics, 21(1), 29-40.
- Ashrafian, H. (2015a). Artificial intelligence and robot responsibilities: Innovating beyond rights. Science and engineering ethics, 21(2), 317-326.
- Athanassoulis, N. (2019). Virtue Ethics. , Virtue Ethics. : IEP
- Audi, R. (2004). The Good in the Right: A Theory of Intuition and Intrinsic Value. : Princeton University Press. Retrieved from https://www.ebook.de/de/product/3982514/robert_audi_the_good_in_the_right.html
- Awad, E. et al. (2018). The Moral Machine experiment. Nature, 563(7729), 59-64. doi:10.1038/s41586-018-0637-6.
- Aziz-Alaoui, M. & Bertelle, C. (2007). Emergent properties in natural and artificial dynamical systems. : Springer Science & Business Media.
- Azoulay, A. (2018). Towards an Ethics of Artificial Intelligence. Retrieved from https://unchronicle.un.org/article/towards-ethics-artificial-intelligence
- Baldwin, T. (2010). In J. Skorupski (Ed.), The Open Question Argument (pp. 286-296). : Routledge.
- Barrett, A. B. (2014). An integration of integrated information theory with fundamental physics. Frontiers in Psychology, 5,. doi:10.3389/fpsyg.2014.00063.
- Baumgaertner Bert & Floridi, L. (2016). Introduction: The philosophy of information. Topoi, 35(1), 157-159.
- Bechor Tamir, Zhang Hengwei & Cruz, L. (2018). Holistic Understanding of Challenges with Autonomous Vehicles. , Holistic Understanding of Challenges with Autonomous Vehicles. :
- Bedau, M. A. (2008). Is weak emergence just in the mind?. Minds and Machines, 18(4), 443-459.
- Bello, P., Licato, J. & Bringsjord, S. (2015). Constraints on freely chosen action for moral robots: consciousness and control. , Constraints on freely chosen action for moral robots: consciousness and control. : IEEE

- Bench-Capon, T. (2020). Ethical approaches and autonomous systems. Artificial Intelligence, 281, 103239. doi:10.1016/j.artint.2020.103239.
- Bench-Capon, T. J., Sartor, G. & others (2000). Using values and theories to resolve disagreement in law. Legal knowledge and information systems: Jurix, , 73-84.
- Bertalanffy, L. v. (1969). General system theory: Foundations, development, applications. :.
- Bezhanishvili, G. & Fussner, W. (2013). An introduction to symbolic logic. Available from the webpage http://www. cs. nmsu. edu/historical-projects.
- Bibleinfo (2020). Ten commandments list | Bibleinfo.com. Retrieved from https://www.bibleinfo.com/en/topics/ten-commandments-list
- Bleske-Rechek, A., Nelson, L. A., Baker, J. P., Remiker, M. W. & Brandt, S. J. (2010). Evolution and the Trolley problem: People save five over one unless the one is young, genetically related, or a romantic partner. , Evolution and the Trolley problem: People save five over one unless the one is young, genetically related, or a romantic partner (Vol. 4). :
- Bongiovanni, G., Postema, G., Rotolo, A., Sartor, G., Valentini, C. & Walton, D. (2018). Handbook of Legal Reasoning and Argumentation. : Springer.
- Bonnemains, V., Saurel, C. & Tessier, C. (2018). Embedded ethics: some technical and ethical challenges. Ethics and Information Technology, 20(1), 41-58.
- Bostrom, N. (2014). Superintelligence. : Oxford University Press. Retrieved from https://www.ebook.de/de/product/21968826/nick_bostrom_superintelligence.html
- Bremermann, H. J. & others (1962). Optimization through evolution and recombination. Self-organizing systems, 93, 106.
- Brennan, A. & Lo, N. (2010). (Ed.), The Environment (pp. 780-792). : Routledge.
- Brennan, A. & Lo, Y.-S. (2016). Environmental Ethics. In E. N. Zalta (Ed.), Environmental Ethics (Winter 2016 ed.). : Metaphysics Research Lab, Stanford University
- Bringsjord, S., Arkoudas, K. & Bello, P. (2006). Toward a general logicist methodology for engineering ethically correct robots. IEEE Intelligent Systems, 21(4), 38-44.
- Brink, D. O. (2006). (Ed.), Some Forms and Limits of Consequentialism (pp.). :.
- Brundage, M. et al. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. arXiv preprint arXiv:1802.07228.
- Buechner, J. (2018). Two New Philosophical Problems for Robo-Ethics. Information, 9(10), 256.
- Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society, 3(1), 2053951715622512.
- Bynum, T. W. (2016). In L. Floridi (Ed.), Informational metaphysics (pp. 203-218). : OUP.
- Campolo, A., Sanfilippo, M., Whittaker, M., Crawford, K. & Selbst, A. (2017). AI Now Report 2017.
- Cao, F., Zhang, J., Song, L., Wang, S., Miao, D. & Peng, J. (2017). Framing Effect in the Trolley Problem and Footbridge Dilemma. Psychological Reports, 120(1), 88-101. doi:10.1177/0033294116685866.
- Caplan, R., Donovan, J., Hanson, L. & Matthews, J. (2018). Algorithmic Accountability: A Primer. Data & Society.
- Capurro, R. (2006). Towards an ontological foundation of information ethics. Ethics and information technology, 8(4), 175-186.
- Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M. & Floridi, L. (2017). Artificial Intelligence and the �?Good Society’: the US, EU, and UK approach. Science and Engineering Ethics, , 1-24.
- Cerullo, M. A. (2015). The Problem with Phi: A Critique of Integrated Information Theory. PLOS Computational Biology, 11(9), e1004286. doi:10.1371/journal.pcbi.1004286.
- Chopra, S. & White, L. F. (2011). A legal theory for autonomous artificial agents. : University of Michigan Press.
- Chrisley, R. (2008). Philosophical foundations of artificial consciousness. Artificial intelligence in Medicine, 44(2), 119-137.
- CLAIRE (2020). CONFEDERATION OF LABORATORIES FOR ARTIFICIAL INTELLIGENCE RESEARCH IN EUROPE. Retrieved from https://claire-ai.org/

- Cochrane, A. (2006). Environmental ethics. , Environmental ethics. : The Internet Encyclopedia of Philosophy (IEP)
- Collingwood, L. (2018). Trust in the machine: the case of Autonomous vehicles. Journal of Information Rights, Policy and Practice, 2(2),.
- European Commission (2018). European Group on Ethics in Science and New Technologies; Statement on Artificial Intelligence, Robotics and Autonomous' Systems. Human reproduction and genetic ethics, (1), 1.
- Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y. & Kramer, M. (2017). Moral Decision Making Frameworks for Artificial Intelligence.. , Moral Decision Making Frameworks for Artificial Intelligence.. :
- Copp, D. (2006). The Oxford handbook of ethical theory. : Oxford University Press.
- Crawford, K., Whittaker, M., Elish, M., Barocas, S., Plasek, A. & Ferryman, K. (2016). The AI Now Report: The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term.
- Creswell, J. W. (2014). Research design: Qualitative, quantitative, and mixed methods approaches. : Sage publications.
- Criado, N., Argente, E., Noriega, P. & Botti, V. (2013). Human-inspired model for norm compliance decision making. Information Sciences, 245, 218-239.
- Cupchik, G. (2001). Constructivist realism: An ontology that encompasses positivist and constructivist approaches to the social sciences. , Constructivist realism: An ontology that encompasses positivist and constructivist approaches to the social sciences (Vol. 2). :
- Dameski, A. (2018). A Comprehensive Ethical Framework for AI Entities: Foundations. In M. Iklé, A. Franz, R. Rzepka & B. Goertzel (Eds.), A Comprehensive Ethical Framework for AI Entities: Foundations. Cham: Springer International Publishing
- Dameski, A. (2020). Avoiding Corporate Armageddon: The need for the establishment of a comprehensive ethical framework for AI & automation in the business world. , Avoiding Corporate Armageddon: The need for the establishment of a comprehensive ethical framework for AI & automation in the business world. : Springer International Publishing
- Danaher, J. (2015). Why AI Doomsayers are like sceptical theists and why it matters. Minds and Machines, 25(3), 231-246.
- Danaher, J. (2016). The threat of algocracy: reality, resistance and accommodation. Philosophy & Technology, 29(3), 245-268.
- Danaher, J. et al. (2017). Algorithmic governance: Developing a research agenda through the power of collective intelligence. Big Data & Society, 4(2), 2053951717726554.
- Danielson, P. (2002). Artificial morality: Virtuous robots for virtual games. : Routledge. Retrieved from https://www.amazon.com/Artificial-Morality-Virtuous-Robots-Virtual-ebook/dp/B000FBFA8G?SubscriptionId=AKIAIOBINVZYXZQZ2U3A&tag=chimbori05-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=B000FBFA8G
- DeLapp, K. M. (2019). Meta-ethics. , Meta-ethics. :
- Delvaux, M. (2016). Draft report with recommendations to the Commission on Civil Law Rules on Robotics. European Parliament Committee on Legal Affairs http://www. europarl. europa. eu/sides/getDoc. do.
- DiCarlo, C. (2016). How to Avoid a Robotic Apocalypse: A Consideration on the Future Developments of AI, Emergent Consciousness, and the Frankenstein Effect. IEEE Technology and Society Magazine, 35(4), 56-61.
- Dictionary.com (2019b). Definition of maximization textbar Dictionary.com. , Definition of maximization textbar Dictionary.com. : Dictionary.com
- Dictionary.com (2019a). Pattern Synonyms, Pattern Antonyms. Retrieved from https://www.thesaurus.com/browse/pattern
- Duke, A. A. & Bègue, L. (2015). The drunk utilitarian: Blood alcohol concentration predicts utilitarian responses in moral dilemmas. Cognition, 134, 121-127. doi:10.1016/j.cognition.2014.09.006.
- Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial intelligence, 77(2), 321-357.
- Durante, M. (2011). The online construction of personal identity through trust and privacy. Information, 2(4), 594-620.

- Dzionek-Kozłowska, J. & Rehman, S. (2019). Career Choices and Moral Choices. Changing Tracks in the Trolley Problem. Studies in Logic, Grammar and Rhetoric, 59(1), 177-189. doi:10.2478/slgr-2019-0036.
- Einar Himma, K. (2007). Foundational issues in information ethics. Library Hi Tech, 25(1), 79-94.
- Ess, C. (2009). Floridi's philosophy of information and information ethics: Current perspectives, future directions. The information society, 25(3), 159-168.
- Statistics Explained (2019). Population and population change statistics - Statistics Explained. Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php/Population_and_population_change_statistics
- Fagella, D. (2019). AI in the Accounting Big Four – Comparing Deloitte, PwC, KPMG, and EY. Retrieved from https://emerj.com/ai-sector-overviews/ai-in-the-accounting-big-four-comparing-deloitte-pwc-kpmg-and-ey/
- Fleischman, W. M. (2015). Just say “no!” to lethal autonomous robotic weapons. Journal of Information, Communication and Ethics in Society, 13(3/4), 299-313.
- Floridi, L. (2002). What is the Philosophy of Information?. Metaphilosophy, 33(1-2), 123-145.
- Floridi, L. (2010). The Cambridge handbook of information and computer ethics. : Cambridge University Press.
- Floridi, L. (2011). The philosophy of information. : OUP Oxford.
- Floridi, L. (2013). The ethics of information. : Oxford University Press.
- Floridi, L. (2016b). In L. Floridi (Ed.), Semantic information (pp. 44-49). : OUP.
- Floridi, L. (2016a). The Routledge handbook of philosophy of information. : Routledge.
- Floridi, L. (2019). The Logic of Information: A Theory of Philosophy as Conceptual Design. : Oxford University Press.
- Floridi, L. et al. (2018). An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations.
- Foot, P. (1967, 2002). The Problem of Abortion and the Doctrine of the Double Effect. Oxford Review, , 19-32. doi:10.1093/0199252866.003.0002.
- Franklin, S. & Graesser, A. (1996). Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents. , Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents. : Springer
- Fultot, M. F. (2016). Ethics of Entropy. APA Newsletter, 15(2), 4-9. Retrieved from https://www.academia.edu/25665067/Ethics_of_Entropy_Full_issue_
- Gabbay, D., Horty, J., Parent, X., van der Meyden, R. & van der Torre, L. (2013). Handbook of deontic logic and normative systems.
- Garrett, J. (2004). A Simple and usable (although incomplete) ethical theory based on the ethics of WD Ross. Retrieved from https://people.wku.edu/jan.garrett/ethics/rossethc.htm
- Gert, B. (2010). In J. Skorupski (Ed.), Hobbes (pp. 88-98). : Routledge.
- von Glasersfeld, E. (2001). An exposition of constructivism: Why some like it radical. , An exposition of constructivism: Why some like it radical. : Springer
- Goff, P. (2009). 6. In D. Skrbina (Ed.), Can the panpsychist get around the combination problem? (pp. 129-135). : John Benjamins Publishing Company.
- Goff, P., Seager, W. & Allen-Hermanson, S. (2017). Panpsychism. In E. N. Zalta (Ed.), Panpsychism (Winter 2017 ed.). : Metaphysics Research Lab, Stanford University
- Goodall, N. (2014a). Ethical decision making during automated vehicle crashes. Transportation Research Record: Journal of the Transportation Research Board, (2424), 58-65.
- Goodall, N. J. (2014b). Machine ethics and automated vehicles. , Machine ethics and automated vehicles. : Springer
- Goodall, N. J. (2016). Away from trolley problems and toward risk management. Applied Artificial Intelligence, 30(8), 810-821.
- Goodall, N. J. (2017). From trolleys to risk: models for ethical autonomous driving.
- Grodzinsky, Miller. & Wolf (2008). The ethics of designing artificial agents.
- Grodzinsky, F. S., Wolf, M. J. & Miller, K. W. (2011). Quantum computing and cloud computing: humans trusting humans via machines. , Quantum computing and cloud computing: humans trusting humans via machines. : IEEE

- Guido Governatori Monica Palmirani, R. R. A. R. & Sartor, G. (2005). Norm Modifications in Defeasible Logic.
- Gunkel, D. J. (2012). The machine question: critical perspectives on AI, robots, and ethics. : MIT Press.
- Gunkel, D. J. (2014). A vindication of the rights of machines. Philosophy & Technology, 27(1), 113-132.
- Gunkel, D. J. (2017). The other question: can and should robots have rights?. Ethics and Information Technology, , 1-13.
- Gunkel, D. J. & Bryson, J. (2014). Introduction to the special issue on machine morality: The machine as moral agent and patient. Philosophy & Technology, 27(1), 5-8.
- Haakonssen, K. (2010). (Ed.), Early modern natural law (pp. 76-87). : Routledge.
- Hagendorff, T. (2019). The ethics of AI ethics: an evaluation of guidelines. arXiv preprint arXiv:1903.03425, 30, 99-120. doi:10.1007/s11023-020-09517-8.
- Hahn, H. (2011). Justifying Feasibility Constraints on Human Rights. Ethical Theory and Moral Practice, 15(2), 143-157. doi:10.1007/s10677-011-9275-x.
- Haken, H. (1983). Synergetics: An Introduction. : Springer.
- Han, P. (2015). Towards a superintelligent notion of the good: Metaethical considerations on rationality and the good, with the singularity in mind (Doctoral Dissertation). Retrieved from
- Harshman, N. L. (2016). In L. Floridi (Ed.), Physics and information (pp.). : OUP.
- Harter, N., Dean, M. & Evanecky, D. (2004). The ethics of systems thinking. , The ethics of systems thinking. : American Society for Engineering Education
- Held, V. (2006). In D. Copp (Ed.), The Ethics of Care (pp.). : Oxford University Press.
- Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent?. Ethics and Information Technology, 11(1), 19-29.
- Hooker, B. (2016). Rule Consequentialism. In E. N. Zalta (Ed.), Rule Consequentialism (Winter 2016 ed.). : Metaphysics Research Lab, Stanford University
- Horgan, J. (2015). Can Integrated Information Theory Explain Consciousness?. Scientific American Blog Network. Retrieved from https://blogs.scientificamerican.com/cross-check/can-integrated-information-theory-explain-consciousness/
- Huebner, B. (2013). Macrocognition: A theory of distributed minds and collective intentionality. : Oxford University Press.
- Hurka, T. (2006). In D. Copp (Ed.), Value Theory (pp.). :.
- Hursthouse, R. & Pettigrove, G. (2018). Virtue Ethics. In E. N. Zalta (Ed.), Virtue Ethics (Winter 2018 ed.). : Metaphysics Research Lab, Stanford University
- Hussain, N. J. Z. (2006). In J. Skorupski (Ed.), Error theory and fictionalism (pp.). : Routledge.
- Johnston, I. (2014). Lecture on Aristotle's Nicomachean Ethics. Retrieved from https://johnstoniatexts.x10host.com/lectures/ethicslecture.htm
- Julia Angwin, J. L. (2016). Machine Bias. ProPublica. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- Kahn Jr, P. H. et al. (2012). Do people hold a humanoid robot morally accountable for the harm it causes?. , Do people hold a humanoid robot morally accountable for the harm it causes?. : ACM
- Kanungo, S. & Jain, V. (2007). General Systems Theory: A Guiding Framework for IS Research. , General Systems Theory: A Guiding Framework for IS Research. :
- Key, T. M., Azab, C. & Clark, T. (2019). Embedded ethics: How complex systems and structures guide ethical outcomes. Business Horizons.
- Killen, M., Smetana, J. G. & Pratt, M. W. (2006). Handbook of Moral Development. Canadian Psychology, 47(4), 344-346.
- Kitto, K. L. & Sylvester, B. (2002). A multidisciplinary approach to teaching ethical considerations in engineering technology. , A multidisciplinary approach to teaching ethical considerations in engineering technology (Vol. 3). : IEEE
- Kleeman, J. H. (2017). How We Can be Justified in Creating a System of Control for Superintelligence (Doctoral Dissertation). Retrieved from

- Klir, G. J. (2001). Facets of Systems Science. 2nd edn., IFSR International Series on Systems Science and Engineering, vol. 15. : Kluwer/Plenum, New York.
- Koch Christof & Tononi, G. (2014). Christof Koch and Giulio Tononi on Consciousness at the FQXi conference 2014 in Vieques. Retrieved from https://www.youtube.com/watch?v=1cO4R_H4Kww
- Lewis, C. S. (2017). Христијанство [Mere Christianity (Macedonian translation)]. : Метанојa.
- Lexico (2019a). Cognition textbar Definition of Cognition by Lexico. Lexico Dictionaries textbar English. Retrieved from https://www.lexico.com/en/definition/cognition
- Lexico (2019b). Reasoning textbar Definition of Cognition by Lexico. Lexico Dictionaries textbar English. Retrieved from https://www.lexico.com/en/definition/reasoning
- Liao, B. (2019). Formal Argumentation (presentation). Journal of Logic and Computation, 29(2), 215-240.
- Liao, B., Slavkovik, M. & van der Torre, L. (2018). Building Jiminy Cricket: An Architecture for Moral Agreements Among Stakeholders. arXiv preprint arXiv:1812.04741.
- Lim, H. C., Stocker, R. & Larkin, H. (2008). Ethical trust and social moral norms simulation: A bio-inspired agent-based modelling approach. , Ethical trust and social moral norms simulation: A bio-inspired agent-based modelling approach. : IEEE Computer Society
- Lin, P., Abney, K. & Bekey, G. A. (2011). Robot ethics: the ethical and social implications of robotics. : MIT press.
- Long, A. (2010). (Ed.), Later ancient ethics (pp. 78-88). : Routledge.
- Luhmann, N.Baecker, D. (Ed.) (2013). Introduction to systems theory. : Polity Cambridge.
- Malerba, A. (2017). Interpretive Interactions among Legal Systems and Argumentation Schemes (Doctoral Dissertation). Retrieved from
- Mansouri, N., Goher, K. & Hosseini, S. E. (2017). Ethical framework of assistive devices: review and reflection. Robotics and biomimetics, 4(1), 19.
- Martin, G. T. (2014). The Ethics and Politics of Holism.
- Martin, K. (2018). Ethical Implications and Accountability of Algorithms. Journal of Business Ethics, , 1-16.
- Mathews, F. (2011). Panpsychism as paradigm. The Mental as Fundamental: New Perspectives on Panpsychism, , 141-155.
- McNaughton, D. & Rawling, P. (2006). In D. Copp (Ed.), Deontology (pp. 424-458). : Oxford University Press.
- Meek, T., Barham, H., Beltaif, N., Kaadoor, A. & Akhter, T. (2016). Managing the ethical and risk implications of rapid advances in artificial intelligence: A literature review. , Managing the ethical and risk implications of rapid advances in artificial intelligence: A literature review. : IEEE
- Meizhen, D. & Zhaoming, G. (2016). Robot Warriors autonomously employing lethal weapons: Can they be morally justified?. , Robot Warriors autonomously employing lethal weapons: Can they be morally justified?. : IEEE
- Merriam-Webster (2019b). Complex | Definition of Complex by Merriam-Webster. Retrieved from https://www.merriam-webster.com/dictionary/complex
- Merriam-Webster (2019a). Cybernetics | Definition of Cybernetics by Merriam-Webster. Retrieved from https://www.merriam-webster.com/dictionary/cybernetics
- Miller, J. G. (2001). Can systems theory generate testable hypotheses? From Talcott Parsons to living systems theory. , Can systems theory generate testable hypotheses? From Talcott Parsons to living systems theory. : Springer
- Mingers, J. (1991). The cognitive theories of Maturana and Varela. Systems Practice, 4(4), 319-338.
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S. & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 2053951716679679.
- Modgil, S. & Prakken, H. (2014). The ASPIC+ framework for structured argumentation: a tutorial. Argument & Computation, 5(1), 31-62.
- Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. IEEE intelligent systems, 21(4), 18-21.
- Morley, J., Floridi, L., Kinsey, L. & Elhalal, A. (2019). From What to How. An Overview of AI Ethics Tools, Methods and Research to Translate Principles into Practices. arXiv preprint arXiv:1905.06876.

- Muehlhauser, L. & Helm, L. (2012). The singularity and machine ethics. , The singularity and machine ethics. : Springer
- Mulgan, G. (2014). True Collective Intelligence? A Sketch of a Possible New Field. Philosophy & Technology, 27(1), 133-142.
- Mørch, H. H. (2018). Is the Integrated Information Theory of Consciousness Compatible with Russellian Panpsychism?. Erkenntnis, , 1-21.
- United Nations (1945). Charter of the United Nations. Retrieved from https://www.un.org/en/charter-united-nations/index.html
- United Nations (1948). Universal Declaration of Human Rights. Retrieved from http://www.un.org/en/universal-declaration-human-rights/index.html
- Neely, E. L. (2013). Machines and the Moral Community. Philosophy & Technology, 27(1), 97-111. doi:10.1007/s13347-013-0114-y.
- Nowak, M., Klok, J., Schwarz, I., Arbour, L. & Johnsson, A. B. (2005). Human rights: handbook for parliamentarians (Vol. 8). : Inter-Parliamentary Union.
- Nugent, P. D. (2018). Punchlines and Plot Twists. Exploring the Relationships between Consciousness, Cybernetics, and Information Theory. , Punchlines and Plot Twists. Exploring the Relationships between Consciousness, Cybernetics, and Information Theory. :
- Nuotio, K. (2010). Systems Theory with Discourse Ethics: Squaring the Circle?: Comment on Marcelo Neve s Zwischen Themis und Leviathan. No foundations: journal of extreme legal positivism, 2010(7), 59-85.
- O'Toole, G. (2015). Simplicity is the Ultimate Sophistication. Retrieved from https://quoteinvestigator.com/2015/04/02/simple/
- Oizumi, M., Albantakis, L. & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. PLoS computational biology, 10(5), e1003588.
- van Oudenhoven, J. P., de Raad, B., Carmona, C., Helbig, A.-K. & van der Linden, M. (2012). Are virtues shaped by national cultures or religions?. Swiss Journal of Psychology.
- Pagallo, U. (2017). When morals ain’t enough: robots, ethics, and the rules of the law. Minds and Machines, 27(4), 625-638.
- Paolo, E. A. D. (2005). Autopoiesis, Adaptivity, Teleology, Agency. Phenomenology and the Cognitive Sciences, 4(4), 429-452. doi:10.1007/s11097-005-9002-y.
- Patrignani, N. & Whitehouse, D. (2014). Slow Tech: a quest for good, clean and fair ICT. Journal of Information, Communication and Ethics in Society, 12(2), 78-92.
- Patrignani, N. & Whitehouse, D. (2015). Slow Tech: a roadmap for a good, clean and fair ICT. Journal of Information, Communication and Ethics in Society, 13(3/4), 268-282.
- Pereira, L. M., Saptawijaya, A. & others (2016). Programming machine ethics (Vol. 26). : Springer. Retrieved from https://www.ebook.de/de/product/25758883/ari_saptawijaya_luis_moniz_pereira_programming_machine_ethics.html
- Peterson, J. (2017). The Jordan B. Peterson podcast - Episode 20 - Ideology, Logos & Belief. Jordan Peterson. Retrieved from https://www.jordanbpeterson.com/podcast/episode-20/
- Piaget, J. (1970). Structuralism. : Basic Books, Inc..
- Podschwadek, F. (2017). Do androids dream of normative endorsement? On the fallibility of artificial moral agents. Artificial Intelligence and Law, 25(3), 325-339.
- Powers, T. M. (2006). Prospects for a Kantian machine. IEEE Intelligent Systems, 21(4), 46-51.
- Prigogine, I. (2001). New perspectives on complexity. , New perspectives on complexity. : Springer
- Quinn, P. L. (2006). In D. Copp (Ed.), Theological Voluntarism (pp.). : Oxford University Press.
- Raafat, R. M., Chater, N. & Frith, C. (2009). Herding in humans. Trends in Cognitive Sciences, 13(10), 420-428. doi:10.1016/j.tics.2009.08.002.
- Rawls, J. Башевски, Д. (Ed.) (2002, 1997, 1971). Теорија на праведноста [A Theory of Justice (Macedonian translation)]. Skopje, Macedonia: Слово.

- Reddy, T. (2017). The code of ethics for AI and chatbots that every brand should follow. Watson. Retrieved from https://www.ibm.com/blogs/watson/2017/10/the-code-of-ethics-for-ai-and-chatbots-that-every-brand-should-follow/
- Richardson, K. (2016). Sex robot matters: slavery, the prostituted, and the rights of machines. IEEE Technology and Society Magazine, 35(2), 46-53.
- Riveret, R., Gao, Y., Governatori, G., Rotolo, A., Pitt, J. & Sartor, G. (2019). A probabilistic argumentation framework for reinforcement learning agents. Autonomous Agents and Multi-Agent Systems, , 1-59.
- Robertson, S. (2006). In J. Skorupski (Ed.), Reasons, Values, and Morality (pp.). : Routledge.
- Rosen, R. (1978). Biology and systems research: an overview. , Biology and systems research: an overview. : Springer
- Ross, W. D.Stratton-Lake, P. (Ed.) (2002). The Right and the Good (British Moral Philosophers). Oxford: Clarendon Press.
- Ryle, G. (2009). The concept of mind. : Routledge.
- Salkind, N. J. (2010). Encyclopedia of research design (Vol. 3). : Sage.
- Sander-Staudt, M. (2011). Care ethics. , Care ethics. : The Internet Encyclopedia of Philosophy (IEP)
- Saptawijaya, A. (2015). Machine ethics via logic programming (Doctoral Dissertation). Retrieved from
- Sayre-McCord, G. (2014). Metaethics. In E. N. Zalta (Ed.), Metaethics (Summer 2014 ed.). : Metaphysics Research Lab, Stanford University
- Schaerer, E., Kelley, R. & Nicolescu, M. (2009). Robots as animals: A framework for liability and responsibility in human-robot interactions. , Robots as animals: A framework for liability and responsibility in human-robot interactions. : IEEE
- Schäffner, V. (2018). Caught Up in Ethical Dilemmas: An Adapted Consequentialist Perspective on Self-Driving Vehicles. Envisioning Robots in Society--Power, Politics, and Public Space: Proceedings of Robophilosophy 2018/TRANSOR 2018, 311, 327.
- Scheutz, M. (2017). The Case for Explicit Ethical Agents.. AI Magazine, 38(4), 57-64.
- Schroeder, M. (2016). Value Theory. In E. N. Zalta (Ed.), Value Theory (Fall 2016 ed.). : Metaphysics Research Lab, Stanford University
- Schulzke, M. (2011). Robots as weapons in just wars. Philosophy & Technology, 24(3), 293.
- Schulzke, M. (2013). Autonomous weapons and distributed responsibility. Philosophy & Technology, 26(2), 203-219.
- Searle, J. R., Tononi, G. & Koch, C. (2013). Can a Photodiode Be Conscious?. Retrieved from https://www.nybooks.com/articles/2013/03/07/can-photodiode-be-conscious/
- Shannon, C. E. (1948). A mathematical theory of communication. Bell system technical journal, 27(3), 379-423.
- Simpson, D. L. (2012). William David Ross (1877-1971). :. Retrieved from https://www.iep.utm.edu/ross-wd/
- Skorupski, J. & others (2010). The Routledge companion to ethics. : Routledge.
- Skrbina, D. (2009a). Mind that abides: panpsychism in the new millennium (Vol. 75). : John Benjamins Publishing.
- Skrbina, D. (2009b). In D. Skrbina (Ed.), Panpsychism in history: an overview (pp.). :.
- Skyttner, L. (2005). Systems theory and the science of military command and control. Kybernetes, 34(7/8), 1240-1260.
- Smith, R. (2017). A neuro-cognitive defense of the unified self. Consciousness and cognition, 48, 21-39.
- Spät, P. (2009). 8. In D. Skrbina (Ed.), Panpsychism, the Big-Bang-Argument, and the dignity of life (pp. 159-176). : John Benjamins Publishing Company.
- Stankovic Mirjana, Gupta Ravi, Rossert Bertrand Andre, Myers Gordon I. & Nicoli, M. (2017). Exploring Legal, Ethical and Policy Implications of AI.
- Steele, K. & Stefánsson, H. O. (2016). Decision Theory. In E. N. Zalta (Ed.), Decision Theory (Winter 2016 ed.). : Metaphysics Research Lab, Stanford University
- Sullins, J. P. (2013). An ethical analysis of the case for robotic weapons arms control. , An ethical analysis of the case for robotic weapons arms control. : IEEE

- Taddeo, M. (2016). In L. Floridi (Ed.), The moral value of information and information ethics (pp. 361-374). : OUP.
- Taddeo, M. (2017). Trusting Digital Technologies Correctly. Minds and Machines, 27(4), 565-568.
- Taddeo, M. & Floridi, L. (2018). How AI can be a force for good. Science, 361(6404), 751-752.
- Tavani, H. T. (2015). Levels of Trust in the Context of Machine Ethics. Philosophy & Technology, 28(1), 75-90.
- Tavani, H. T. (2018). Can Social Robots Qualify for Moral Consideration? Reframing the Question about Robot Rights. Information, 9(4), 73.
- Thekkilakattil, A. & Dodig-Crnkovic, G. (2015). Ethics aspects of embedded and cyber-physical systems. , Ethics aspects of embedded and cyber-physical systems (Vol. 2). : IEEE
- Thomson, J. J. (1985). The Trolley Problem. The Yale Law Journal, 94(6), 1395. doi:10.2307/796133.
- Tononi, G. (2004). An information integration theory of consciousness. BMC Neuroscience, 5(1), 42. doi:10.1186/1471-2202-5-42.
- Tononi, G. (2015). Integrated information theory. Scholarpedia, 10(1), 4164. doi:10.4249/scholarpedia.4164. Retrieved from http://www.scholarpedia.org/article/Integrated_information_theory
- Tononi, G., Boly, M., Massimini, M. & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. Nature Reviews Neuroscience, 17(7), 450.
- Torrance, S. (2008). Ethics and consciousness in artificial agents. Ai & Society, 22(4), 495-521.
- Torrance, S. (2014). Artificial consciousness and artificial ethics: Between realism and social relationism. Philosophy & Technology, 27(1), 9-29.
- Tzafestas, S. G. (2016). Roboethics. : Springer. Retrieved from https://www.ebook.de/de/product/24866947/spyros_tzafestas_roboethics.html
- Valentinov, V., Hielscher, S. & Pies, I. (2016). Emergence: a systems theory's challenge to ethics. Systemic Practice and Action Research, 29(6), 597-610.
- Van den Hoven, J., Vermaas, P. & Van de Poel, I. (2015). Handbook of ethics, values and technological design. : Springer.
- Van Leeuwen, J. (2014). On Floridi's method of levels of abstraction. Minds and Machines, 24(1), 5-17.
- Varela, F. G., Maturana, H. R. & Uribe, R. (2001). Autopoiesis: the organization of living systems, its characterization and a model. , Autopoiesis: the organization of living systems, its characterization and a model. : Springer
- Von Foerster, H. (2003). Ethics and second-order cybernetics. , Ethics and second-order cybernetics. : Springer
- Wachter, S., Mittelstadt, B. & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics. , 2, eaan6080. doi:10.1126/scirobotics.aan6080.
- Wallach, W. (2010). Robot minds and human ethics: the need for a comprehensive model of moral decision making. Ethics and Information Technology, 12(3), 243-250.
- Wallach, W., Franklin, S. & Allen, C. (2010). A conceptual and computational model of moral decision making in human and artificial agents. Topics in cognitive science, 2(3), 454-485.
- Walton, D. (2009). Argumentation theory: A very short introduction. , Argumentation theory: A very short introduction. : Springer
- Wenar, L. (2015). Rights. In E. N. Zalta (Ed.), Rights (Fall 2015 ed.). : Metaphysics Research Lab, Stanford University
- Wiltshire, T. J. (2015). A prospective framework for the design of ideal artificial moral agents: Insights from the science of heroism in humans. Minds and Machines, 25(1), 57-71.
- Zadeh, L. A. (2001). The role of fuzzy logic in modeling, identification and control. , The role of fuzzy logic in modeling, identification and control. : Springer
- Müller, V. C. (Ed.) (2016). Fundamental issues of artificial intelligence (Vol. 376). : Springer. Retrieved from https://www.ebook.de/de/product/27955774/fundamental_issues_of_artificial_intelligence.html
- (2019). International Journal of Ethics and Systems. Emerald | International journal of ethics and systems information. Retrieved from http://www.emeraldgrouppublishing.com/products/journals/journals.htm?id=IJOES

Appendix I. Definitions of Terms and Abbreviations

The following definitions and abbreviations are given and apply solely *for the purpose of this thesis text*. I am not trying to dismiss the philosophical, scientific and linguistic debates that surround many of the concepts provided here, and thus these definitions cannot be considered as final and set.

- **Achievement of Personal Goals (APG)** (Imperative) – the *Achievement of Personal Goals (APG)* is one of the two ethical imperatives that comprise the striving towards a desired *Quality of Life* (see below) of a system.
- **AI entities –** are all entities which are based on artificial intelligence, which means that they have at least some non-biological reasoning capacity embedded in them. AI entities are a subset of the class of 'artificial entities' (see below).
- **Agent** (ethics) – an agent from an ethical perspective is an entity that produces a moral process in a moral scenario.
- **Algorithm** – an algorithm is "... understood intuitively as a set of instructions, expressed in some language, for executing a sequence of operations for solving a problem of some specific type" (Klir, 2001). An algorithm is also defined as "a mathematical construct with 'a finite, abstract, effective, compound control structure, imperatively given, accomplishing a given purpose under given provisions.' [...] algorithms must be implemented and executed to take action and have effects" (Hill, 2015; in Mittelstadt et al. (2016)).
- **Artificial entities –** are the class of entities in the universe which are created by humans and our tools, or by extension, by other artificial entities themselves. This would include AI entities (see above) such as algorithms, computers, robots, autonomous vehicles; but also institutions, states, communities, products, buildings, and similar. However, the distinction between 'artificial' and 'natural' is blurry[67], to say the least (see natural entities below), and used solely for the purpose of this text.
- **Axiom** (ethics) – an ethical axiom is a fundamental, initial ethical and self-evident principle. Axioms are further used to reason ethically in context and build moral calculus (see below).
- **Bad** (moral) – a moral bad is the causation of negative morally-burdened effects—the introduction of unwanted entropy somewhere in the universe (i.e. to the structure of a system). If such causation is done intentionally, it is classified as moral evil (see below).
- **Being** – what is defined as 'Being' is existence of something (e.g. an entity, a relation) in the universe, as opposed to non-existence. That something that exists has its own attributes, a particular place in space and time, and is unique in the sense that no other thing or a relation can exist with the exact same attributes at the exact place in space and time.
- **BDI** - Belief, Desire, Intention (abbreviation), used in studies of multi-agent systems as a collection of some important attributes of those systems.
- **Calculus** (moral) – a moral calculus is a formal and explicit way of moral reasoning that enables making decisions and deciding upon a course of (in)action in moral scenarios.
- **Cognition** – is the mental capacity of an entity to be able to be aware of phenomena and to process them into its internal states. Or, as defined by the Collins Dictionary, cognition is: "1. the mental act or process by which knowledge is acquired, including perception, intuition, and reasoning; 2. the knowledge that results from such an act or process" (Collins English Dictionary, 2019).

[67] And also seems unduly biased towards anthropocentrism.

- **Consciousness** – consciousness is the cognitive capacity of an entity that enables it to be aware of phenomena in the universe. These phenomena are typically accessible in an indirect manner (i.e. through sensory informational inputs), but can also sometimes be accessed directly (i.e. awareness of one's own thoughts, which is self-consciousness). This is not to be confused with *self*-consciousness.
- **Conservation of Personal Continuum (CPC)**, Imperative – *the Conservation of Personal Continuum (CPC)* is one of the two ethical imperatives that comprise the striving towards a desired *Quality of Life* (see below) of a system.
- **Essence**, systemic (see *Structure* below)
- **Entity** – an entity is something that exists (see *Being* above) in the universe as part of it, and as something unique and different from everything else. It functions as a system (see below).
- **Entropy** (metaphysics) **–** Indirectly related to Shannon's notion of thermodynamic entropy, entropy in the sense used in this text (the metaphysical sense) is the cessation or nonexistence of order and thus pattern at a particular place and point in time. Entropy is a natural phenomenon of the universe, but can also be introduced by systems in an intentional or unintentional manner (see bad and evil).
- **Ethics** – in contrast to *morality* (see below) which is the practical side of the coin, ethics is the theoretical. Ethics either:
    - deals with the study of the various moralities that appear in practice,
    - or attempts to derive a unifying theoretical framework that can explain them uniformly,
    - or attempts to design a 'perfect' framework of behavior-defining and modifying rules, concepts and principles (that usually, but not exclusively, fall into the interpersonal domain).
- **Ethics of Systems** (abbreviated: EoS) – when used in uppercase, like here, it refers to the (meta)ethical Framework and its Interface, both which are the result of the work in this thesis.
- **Evil** (moral) – moral evil is the intentional and conscious causation of negative morally-burdened effects (introduction of destructive metaphysical entropy; see bad above).
- **Framework** (ethics) – an (ethical) framework is an organized system of ethical axioms, principles, and methods of representation and reasoning, applicable in moral scenarios.
- **Imperative** (ethics) – an imperative from an ethical perspective is an ultimate systemic goal that is desired and pursued by the entity, and its level of achievement directly influences the level of achievement of that system's Quality of Life (see QoL, APG, CPC). Imperatives are explicit or implicit goals—either for the entity itself, or for its designer, owner or user. An ethical imperative is also a systemic imperative (see below).
- **Imperative** (systemic) – a systemic imperative is an ultimate systemic goal that emerges out of the essence of a system, and is the initiatory causation element of (in)action of an entity in a particular context.
- **Information** – information is understood as pattern in a substrate, that a reasoning entity can 'construct' or 'extract' as an emergent immaterial phenomenon. The substrate can be both material, and immaterial (in the form of other information and data).
- **Instrument** (ethics, systemic) – an instrument is an implicit 'temporary' goal or a resource, which is used to pursue explicit goals in general, who also include imperatives (explicit ultimate goals).
- **Integrated information –** as integrated information is taken the *particular* and unique structure (see below) and attributes of a particular system set S, comprised of particular subset of things T and subset of relations R. If this system set S is modified in a substantial manner, the integrated information within it either changes or is destroyed. Or, in the words of the main authors in this domain, "the irreducibility of a conceptual structure is measured as integrated information" (Tononi et al., 2016).
- **Moral** (see *Morality* below) – when used as a *classifier*, 'moral' holds the meaning of 'something in the universe that belongs to the set of things which pertain to morality', a *morally-relevant* phenomenon,

relation or thing e.g. a moral agent, a moral process, a moral system. When used as a *qualifier*, 'moral' holds the meaning of a positive morality e.g. a moral action that has positive morally-burdened effect(s), an entity that has a favorable moral status, etc.

- **Morality** – in contrast to ethics (see above) which is the theoretical side of the coin, morality is the practical. It is the actual internalized behavior-defining and -modifying rules, concepts and principles that actual past or present people and other entities have taken upon as applicable for themselves or for others.
- **Natural entities –** are the class of entities that 'naturally' and 'spontaneously' 'appear(ed)' in the universe, without human intervention. However, this distinction is significantly contentious, since one can argue that, as humans are the result of nature (are natural entities), whatever we create is also natural (including what we call 'artificial entities', see above). This class would include entities such as ecosystems, animal and plants, mountains, planets, and the whole universe.
- **Patient** (ethics) – a patient from an ethical perspective is an entity that receives morally-burdened effects coming from a moral process in a moral scenario.
- **Quality of Life (QoL)** – Quality of Life is a measure for a desired state that an entity wants to be achieved for itself and the world i.e. the highest quality of life. This measure reflects the level of achievement of the two ethical imperatives: CPC and APG (see above).
- **Scenario** (moral) – a moral scenario is a situation in the universe in which there is at least one *moral agent*, one *moral patient*, at least one *moral process*.
- **System** – a system is something in the universe that exists and is different than everything else. It is a set S comprised of two subsets, T (things) and R (relations). A system and an *entity* (see above) are synonymous for the purpose of this text[68].
- **Structure** (systemic) – every system has a structure. This structure is the particular way in which the set of things (T) and the set of relations (R) are arranged to give the form and function of that particular system. The structure is (part of) the essence of the system, and changes in it reflect as changes in the essence. Typically, a systemic structure guides the patterns of input, processing and output of matter, energy and information.

[68] Without disregarding the debate between systems constructivism versus systems realism, which is briefly discussed in *Chapter III. Towards Ethics of Systems (the Metaethics)*.

Appendix II. Key concepts

This section holds the identified key ethical concepts that the Framework will have to be able to represent in an explicit manner. They are included in the following table, in an alphabetic order.

Table 17: Identified key concepts throughout the literature review

Identified key concept	Definition
Autonomy	The power to decide (whether to decide) (Floridi et al., 2018; p. 12).
Bad, the	Moral bad is comprised of all decisions, actions, consequences, and states of matter that are morally negative i.e. those *things* that devalue, from a moral point of view, regardless of intentionality. When moral bad is caused intentionally, it falls under the subset of moral *evil* (see below). When moral bad is overwhelming, it sometimes is called as a moral *tragedy*. It is contrasted with *the Good* (see below).
Beneficence	Doing what is good or right from a moral perspective. It typically is understood as promoting well-being, preserving dignity, and sustaining the environment and the planet (Floridi et al., 2018; p. 10).
Care	Care is a relationship whereby an entity spends personal resources to preserve and promote the well-being of another entity.
Common-sense morality	"… distinguishes between the obligatory, the permissible, and the supererogatory" (Brink, 2006; p. 385) (see *duty (moral, legal)*, *permission (moral, legal)*, *supererogation* in this Key Concepts section).
Duty (moral, legal)	Duties are commitments and expectations for entities in the moral and legal realms, respectively. These can be requirements, prohibitions, and permissions.
Evil, the	Moral Evil is the intentional and conscious causation of moral bad.
Goals	Goals are states of matters in a global (of the world/Universe) and local sense, that an entity would like to see become reality. They usually, but not always, require active or passive, and direct or indirect targeted participation of the entity itself to be achieved.
Good, the	The Good is typically understood as what is valuable from the perspective of ethics and morality. Ethical theories commonly dictate how the Good can be pursued, maximized, satisficed, improved, respected, protected, conserved, and the like. The methods to do this differ according to the consulted ethical theory. Similar to moral *Right* (see below).
Information	Information is well-formed, meaningful, and veridical data.
Integration / Integrated	Integration is a process by which a set of parts of the world get intertwined, interrelated and interdependent into a new whole—an entity (system). This implies that at least part of their destiny (future behavior and status) depends on the other components of the system, and on the system itself. Their nature ceases to be fully reducible to the nature of the individual components.
Level of Abstraction	A Level of Abstraction is a set of observables (which are interpreted typed variables) about a system. Its purpose is to create a model of a system under study, and track its changes, development, or anything else of interest.
Maleficence	Bringing about decisions, actions, and consequences that are against the Good, the Right, or the good life; and towards the moral bad and evil. Contrasted with *non-maleficence* (see below).

Maximization	“To increase to the greatest possible amount or degree” (Dictionary.com, 2019b). When talking about decision making, it is about finding and pursuing the decision that is the best out of all available ones at a current point of time in a particular scenario. *Mutatis mutandis* for moral decision making. Contrasted with *satisficing*.
Moral irrelevance / indifference	Morally-irrelevant state of matters is such that cannot be qualified as morally-relevant i.e. by using ethical and moral qualifiers: good, bad, evil, right, wrong… Likewise for morally-irrelevant (in)actions. Such states of matters and (in)actions typically, but not always, bear no effect on moral states of matters.
Moral neutrality	Morally-neutral state of matters is one that does not affect (the global status of) the Good or the Right in a positive or a negative manner (or sometimes it evens out). Likewise with a morally-neutral (in)action. This is different from *moral irrelevance* (see below).
Non-maleficence	The opposite of *maleficence* (see above). It is abstaining and refraining (passive) from, or an effort (active) against, causing moral bad (harm).
Obligation (moral, legal)	See *duty* above.
Optimization	See *maximization* above.
Right (moral, legal)	“Rights are entitlements (not) to perform certain actions, or (not) to be in certain states; or entitlements that others (not) perform certain actions or (not) be in certain states” (Wenar, 2015).
Right, the	What is right from a moral perspective is that which complies with moral *duties* (see above) and obligations. It is a deontological qualifier.
Satisficing	Doing what is good enough. When speaking about morality, satisficing means performing an action that promotes the Good above a certain threshold of value, but does not necessarily maximize the Good out of all possible actions that an entity can take (i.e. it leaves the world at a morally better state, but not the best possible) (Brink, 2006; p. 384). Contrasted with *optimization*.
Supererogation	Going beyond the call of duty. Supererogatory actions are understood as good, but not obligatory (McNaughton & Rawling, 2006; p. 426); or bad, but not (morally) wrong (Quinn, 2006; p. 71). The moral entity pursuing these actions, especially at personal cost, is worthy of moral praise up to a certain point (after which such pursuit might be seen as foolish and uncalled for sacrifice).
Wrong, the	The opposite deontological qualifier of *Right* (see above).

Appendix III. Further commentary on Ethics of Systems

1.1.1 Is EoS (too) conservative?

If we judge solely by the four ethical principles of EoS discussed in 4.2.1.2 Ethics of Systems' four ethical principles it might seem that EoS is too conservative. It might seem that it is 'obsessed' with destructive entropy and its removal or avoidance, and only then is concerned with flourishing. This is a faulty perception.

We should recall that (destructive) entropy is an ever-present feature of the universe and the environment (see 3.3.3 Injury and destruction of Being). Systems fight a never-ending and ultimately futile battle against it. This applies even for systems for which the Conservation of Personal Continuum has an implicit, instrumental value. Before any system can be able to pursue any goals, it or other systems (e.g. humans that deploy it) need to ensure that it can retain integrity at least until its goals are achieved.

This is the reason why the battle against destructive entropy gets a temporal 'priority' when applying the ethical principles. A Being has to be conserved (kept integrated) before it can go on and pursue objectives. This cannot be considered as a 'too' conservative approach. However, it might be considered as conservative-enough one.

1.1.2 Is EoS too reliant on heuristics?

I have discussed in Chapter III. Towards Ethics of Systems (the Metaethics), 2.1.6 Method of Levels of Abstraction and 3.4.2 Complexity, that complexity poses an inherent limitation of computational / cognitive capacity for any system (the Bremermann's computational limit of 2 x 10^{47} bits per gram of mass). In everyday life, however, we are dealing with systems whose computational or cognitive power does not even begin to approach this number anyway. This translates into making the use of cognitive shortcuts (e.g. heuristics) be guaranteed.

Human moral reasoning is not only forced to rely on heuristics, but even recognizes it implicitly and makes the best use of it. The purpose of established moral rules and laws is exactly to offer a simplified way to organize processes in a wide variety of moral scenarios in order to avoid a great amount of cognitive overhead. At times even this is not enough so human communities use other techniques, such as delegation and specialization, hierarchical organization, conservative approaches and other to deal with increasingly complex moral scenarios.

For example, the purpose of having specialized roles (e.g. leader of a community, a priest, a sheriff, a mason, a doctor, a knight etc.) is to delegate the management of differing moral scenarios where specialists can deliver better results, in contrast to expecting every person of the community to be skillful in everything. Similarly, hierarchies (e.g. prime minister, ministers, institutions, clerks etc.) provide a level-like partition of scenarios (and hence categories of information) to avoid cognitive overload, among other things. Conservative approaches (e.g. conservative politics) are similarly efficient at lowering complexity by vouching for *a priori* avoidance of a plethora of undesirable situations that have the potential to increase destructive entropy outside of manageable bounds. None of these methods are sound and thorough, but are highly pragmatic for most cases.

EoS in itself is 'reliant' on heuristics because it recognizes that their utilization is inevitable. If we are to model and manage realistic moral scenarios we ought to recognize their key role in moral reasoning and make the best of them, while trying to improve the results they deliver. Of course, this does not mean that EoS is not a complete, coherent framework. Although there remains still a significant effort to improve and further develop the framework, for now it provides sound methodology to analyze and manage moral scenarios that include a variety of entities (including AI entities).

In order to demonstrate this claim, in Chapter IV. Applying Ethics of Systems Framework to AI moral scenarios (the Ethics) I explore a variety of moral scenarios that will utilize different moral theories, and see what results they deliver.

1.1.3 Some properties of Ethics of Systems

1.1.3.1 Fitting Attitudes, agent-relative and agent-neutral ethics, and Universalizable egoism

Ethics of Systems is able to accommodate both agent-relative (Schroeder, 2016; sect. 3.3.) and agent-neutral moral theories. It can do this because it recognizes the so-called *Fitting Attitudes* account of value. Schroeder defines it thus:

> “if the good is what ought to be desired, then there will be two kinds of good. What ought to be desired by everyone will be the “agent neutral” good, and what ought to be desired by some particular person will be the good relative-to that person” (Schroeder, 2016; sect. 3.3.3.).

First, let’s cover agent-relative morality. Every system’s Being has intrinsic value; and this intrinsic value is appreciated primarily by each system itself for itself. This is the view of (moral) egoism, and is the primary sense in which life is valuable. Life can be valuable in a secondary sense also, if systems recognize the value of existence of another system or class of systems, just because *these others* recognize their own value themselves. This is the basis of intersystemic morality that extends beyond pure moral egoism, but adds the *universalizable* in the agent-relative version of *universalizable egoism* (Schroeder, 2016; sect. 3.1.3.). Life can be valuable in an additional, third sense, whereby a system can recognize the value of existence of all (classes of) systems simply because they exist, regardless if they autorecognize this value themselves or not (just like EoS does).

Now let’s turn to agent-neutral morality. Remember that I started the second sentence of the previous paragraph with the claim that every system’s Being has intrinsic value. If a macroethics or a holoethics recognizes this intrinsic value, even without ‘waiting’ for any moral entity to accept it, we are talking about an agent-neutral moral theory (ethics). And if an agent-neutral theory appreciates each system’s own appreciation for its own value, this is the agent-neutral version of universalizable egoism.

1.1.3.2 Computational representability

The foundations of EoS are computationally representable. Even though in this work I will not dive into complex calculus (except where necessary), I aim to demonstrate this claim with the work in Chapter IV. Applying Ethics of Systems Framework to AI moral scenarios (the Ethics).

We have seen that by using the method of LoAs we can conceptualize the various components and properties of moral scenarios, moral entities, moral processes, moral theories, and other miscellaneous (e.g. amoral) phenomena and properties. All the elements of the Framework can be combined in a unified logico-mathematical calculus, which in turns enables translation in programming and ethico-legal language in order to be executed by different participants in moral scenarios.

Some of these are relatively easy to translate into calculus e.g. QoL, APG, CPC, resources, time, space, etc. Others might prove very difficult for such treatment, and might necessarily be treated as self-reported or imprecise e.g. emotional states, instincts, general entropy, etc. Regardless of the aforementioned difficulty, any phenomena of the world that can be anyhow perceived can be potentially represented with observables.

Appendix IV. Attributes of an ethical framework for AI

With the previous development on EoS we can now draw some fundamental attributes that any comprehensive ethical framework for AI entities ought to have.

Table 18: Attributes of an ethical framework for AI

Attribute	Explanation
Foundational	The framework should be axiomatic (that is, it is set up as a system of axioms that can be informationally, logically and computationally represented) and hence necessarily fundamental.
Coherent	The axiomatic system is able to be informationally, logically and computationally expanded to provide solutions to arising ethical problems in context, without issues of incoherence taking place.
Hybrid, multidisciplinary and holistic	The axiomatic base of the framework is conceived with a holistic approach in mind. It draws on existing advances in ethics in general. It also draws on other, ‘non-ethical’ and metaethical disciplines that can help provide more holistic approach, and thus more comprehensive one.
Unified / unifying	The framework should have universalist pretension i.e. it should attempt to unify all the major ethical theories into a single axiomatic system; or at least provide the means to represent them separately or simultaneously. This also relates to human rights, which are in their conception ‘universal’ i.e. attributed to everyone. Universal human rights (which are widely accepted by humanity), alongside major ethical theories, and new advancements in the field of ethics should form the integrated basis of the framework.
Contextual	The framework, when applied, should be able to ‘live in context’, acquire new and modify existing moral knowledge, and adjust to new environment and circumstances. In this sense, it should be able to satisfy the interests of all involved stakeholders, such as business(es), academia, government(s), policy-makers, and the public. It should provide a straightforward methodology for ensuring moral responsibility and accountability on the part of the aforementioned parties.
Applicable to AI entities, and their interaction with the environment	i.e. other AI systems and other systems in general, the world, humans and their systems, animals, legal, financial and social systems, enterprises/business entities, government, etc.
Translatable and implementable through engineering, internal policy, and legal tools	The framework should be implementable through engineering, programming, and project-building practices and activities. Additionally, it should also be implementable in internal ethical codices, management styles and HR practices and activities, and similar. Finally, the framework or suitable parts of it should be easily codifiable into law by policy-makers—and vice-versa. All the above should be performed in consultation with all involved stakeholders and their representatives to ensure sustainable and stable legal solutions.

Ethics of Systems satisfies the above attributes *in principle*. Concrete implementations might choose to disregard or abstract certain components, but this might result in biased or incoherent effects. As you can see, the above draw upon systems theory to inform the basic structure of an ethical framework (in this case for AI entities). The attributes have already been included in Dameski (2018) and Dameski (2020).