RADBOUD UNIVERSITY

# What We Can Do Now About Superintelligence

RECONCILING TENSIONS AROUND OPENNESS IN AI DEVELOPMENT VIA PATH DEPENDENCE REASONING

BACHELOR'S THESIS IN ARTIFICIAL INTELLIGENCE


*Author:*
Max REDDEL

*Student Number:*
s4830709

*Supervisor:*
Dr. W.F.G. HASELAGER (PIM)
Associate Principal investigator - Donders Centre for Cognition
Associate Professor - Artificial Intelligence

*Second reader:*
Dr. J.H.P. KWISTHOUT (JOHAN)
Assistant professor - Artificial Intelligence
Associate Principal Investigator - Donders Centre for Cognition
Associate Principal Investigator - Donders Institute for Brain,
Cognition and Behaviour



**Abstract**

We can observe a clash between people that prioritize near-term issues and people that prioritize long-term issues in AI. We show that both factions are reconcilable. For this purpose, we use the long-term issue of superintelligence and the near-term issue of openness in AI development. We demonstrate how openness can be charted as a multi-dimensional space and traversing through this space can be modelled as a dynamic process. This process turns out to be equilibrium dependent with a unique lock-in in the most open corner of the openness space. Avoiding this lock-in is desirable for both factions. Concretely, we recommend to go for maximal openness in safety techniques, organizational goals and composition of ethics boards. Minimal openness should be maintained in science, source code, data and development tools. Further near-term AI issues might also exhibit undesirable lock-ins.


11th of January 2020

# Contents

# List of Figures

## List of Tables

# 1 Introduction

## 1.1 Background

Technology brings along both benefits and risks for society. Artificial Intelligence (AI) is an especially powerful technology that can consequently bring along fairly powerful benefits and potentially devastating consequences. Much discussion has been taking place around such risks. When zooming in on these discussions, one cannot miss to identify (at least) two kinds of concerns – one that focuses more on near-term AI risks while the other one is more oriented towards the long-term risks. Both kinds of concerns address the enablement of safe AI development, deployment and governance in order to minimize risks and maximize the net positive outcome for society. The near-term concerns point towards a plethora of topics like cyber-attacks, autonomous weapons, technological labor displacement, algorithmic bias, deepfakes and autonomous vehicles.[1] The long-term concerns include foremost artificial superintelligence (ASI) and its potential as an existential threat to humanity.[2]

**Definition 1.1: Artificial Superintelligence**

An ASI can be regarded as an agent that exceptionally outperforms any human's cognitive performance in any possible domain [2]. This would imply that an ASI is not only better in arithmetic or calculus than an average human (or even the most intelligent human), but the ASI is also better in any other possible human domain such as social manipulation, chemistry or arts. Several authors have sketched how ASI can be considered as an existential threat to humanity. Introducing these sketches would go beyond the scope of this paper. For thorough elaborations, see [3]–[8].

At first glance, one could be feel relief given this situation: There are thinkers that worry about the near-term issues and there are people that worry about the long-term issues. Thus, both important kinds of issues are taken care of. However, not all of the concerned parties seem to see eye to eye.

For instance, *AI Now Institute* co-founder Kate Crawford claims that the long-termists have grabbed a lot of media attention and funding. She is referring to attention and funding that has been taken *away* from "the very real problems with artificial intelligence today" [9]. Furthermore, one of the leading AI experts of our time – Andrew Ng – famously proclaimed that "the reason [he doesn't] worry about AI turning evil is the same reason [he doesn't] worry about overpopulation on Mars." [10]. Ng implies that the problem is still far off into the future and that the hype over killer robots and ASI is an "unnecessary distraction" from urgent issues like labor displacement [11], [12]. The famous roboticist Rodney Brooks jumps on the same bandwagon by dismissing the long-term concerns due to timeline reasons.[3] According to Brooks, artificial general intelligence (AGI) can realistically be 100 to 500 years off while current AI risks are very real and urgent [13].

**Definition 1.2: Artificial General Intelligence**

In the framework of this paper, we consider an AGI as an artificially constructed agent which has the capability of goal optimization over a wide range of (human) domains. Furthermore, such an agent should be able to learn in a goal-directed, autonomous and adaptive way [14]. The performance of an AGI should be comparable to an average human's performance. For this purpose, some researchers and philosophers prefer to grant an agent the status of exhibiting general intelligence by passing a test. Honorable mentions are the Turing test, the Wozniak test, the robot college student test and the employment test. For more information on these tests, see for example [15].

Some proponents of the long-term minded subgroup seem to raise similar concerns into the other direction. The long-termists consider the existential risk, which is posed by ASI, as bigger in scope and more neglected than near-term issues [16]. In their view, the vast number of future lives overshadows the current number of living humans and thus the long-term future should be clearly prioritized [17]. Resources like (media) attention, funding and talent are generally restricted and not everybody agrees how these resources should be distributed to achieve the most good [18].

Mitigating one kind of risk rather than the other by allocating the available resources to the corresponding risk might seem like a zero-sum game to some. However, it is worth investigating this perception more thoroughly as it could also simply be the case that overly-certain and loud individuals hijack the debate and fail to represent the true disagreements. It would not be a zero-sum game between these two kind of foci if there would be some overlap in their goals and thus potential for synergy. The classic comedy *Monty Python's Life of Brian* [19] might come to my mind where the *People's Front of Judea* and the *Judean People's Front*, who both hate and oppose the Romans, focus on their differences and waste their resources to fight each other instead of pursuing their common goal of breaking free from the Roman occupation of Judea. Since their goals are hugely overlapping, their quarrel is preposterous to the movie viewer.

[1] This list is not an exhaustive one. However, these topics are some of the currently salient ones in late 2019.

[2] This is not the only long-term concern. Some researchers argue that e.g. labor displacement is also regarded as a potential long-term concern [1].

[3] Timelines and take-off times have been discussed in length by e.g. Nick Bostrom in his 2014 book (chapter 4) [3] and have a role in section 4.4 of this paper.

## 1.2 Research Questions

An interesting question would be whether the near-termists and long-termists are behaving similar to the two factions from *Monty Python's Life of Brian*. Is the situation really a zero-sum game as the louder speakers seem to indicate? Is there potentially a common ground between near-termists and long-termists that could be focused on instead of the differences? There are several angles to approach this debate. And some researchers like Stephen Cave, Seán Ó hÉigeartaigh [1], Seth Baum [20] and Jess Whittlestone [21] are already trying to disentangle this debate and reconcile the two factions. This paper is an attempt to contribute to this endeavor. Therefore, our first research question addresses a potential reconciliation. The matter becomes more concrete if we use some specific near-term issue such as openness in AI development. In consequence, our first research question is the following one:

**First Research Question**

Can a near-term focus and a long-term focus be reconciled by using an early long-term perspective on shaping openness in AI development?

Let us unpack *openness in AI development*.[4] An elaborate definition can be found in section 3. Openness can refer to how open AI engineers and researchers are when it comes to making aspects of their research publicly available. One could argue to keep specific aspects of AI development as open as they are, to make some of them more open and some less open. A list of all aspects of openness are discussed in section 3.1.

Let us assume that reconciliation of the two factions is indeed possible. We can use our analysis to infer specific action recommendations with respect to openness. Therefore, our second research questions can be formulated as:

**Second Research Question**

Given that a reconciliation is possible, what action recommendations can be given to researchers and engineers with respect to openness in AI development?

## 1.3 Methodology

In order to answer the first research question, we created a semi-formal argument. This method implies a careful check of the soundness of the argument and some way to show that the assumptions hold. The potential falsification of this paper's results are simplified due to the explicitness of this method. The soundness can be checked fairly easy as the argument contains only simple propositions, conjunctions and implications. Furthermore, the intermediate assumptions can be viewed in isolation. Given questionable aspects of our argument, the reader should be enabled to point out a flaw in the argument's structure or show that at least one of the assumptions does not hold. Like this, healthy debate should be made easy. Showing that both parties are reconcilable, implies in our case to show how to reconcile them. Thence, we use this information to answer the second research question. Furthermore, we have to create a multi-dimensional space of openness and model the traversing of this space as a dynamic process which uses path dependence theory. With the methods being sketched, let us turn to the outline of this paper which will provide more details.

## 1.4 Outline

The main argument shall be known as the *Reconciliation Argument* which will be introduced in section 2. To show that *the* critical assumption of the Reconciliation Argument is true, two supporting informal arguments will be introduced in section 5 and 6 − which we call the *Near-Term Benefit Argument* and *Long-Term Benefit Argument*, respectively.

In order to understand these two arguments, the so-called *openness space* has to be charted (see section 3). First, we elaborate on the eleven dimensions of openness in section 3.1. Then, we explore the expected societal value in two flavors (see section 3.2). Both, the Near-Term and the Long-Term Benefit Argument require an elaboration openness and its potential impacts. Both arguments also require the concept of *Path Dependence* which is explained in full length in section 4. To provide some intuition, the following shall be sufficient: Path dependence is strongly connected to lock-in situations where reversing decisions is severely impeded or even made impossible [22]. In the corresponding section, we elaborate on some standard examples (see section 4.1), the theory behind path dependence (see section 4.2), how to apply path dependence theory on openness (see section 4.3) and why it is hard to escape a lock-in situation in the openness space at a late point in time (see section 4.4).

Having introduced the openness space and path dependence, the Near-Term and Long-Term Benefit Argument can be unfolded. Finally, the conclusion in section 7 provides a definite answer to the two research questions. Furthermore, we infer general lessons from applying path dependence reasoning on AI issues which are pressing in the near-term.

## 1.5 Target audience

Let us assume that we would be able to show with this approach that near-termists and long-termists can be reconciled. Who could benefit from this knowledge? Is the result of this paper actionable?

The target group of this paper is not the law-maker, not the politician and not the public. It is addressed foremost at AI engineers, AI researchers in both academia and in

[4] To improve the reading flow in the remainder of this paper, we refer to *openness in AI development* simply as *openness*.

the industry and decision makers in the corporate tech domain. The knowledge *how* to reconcile the near-termists and the long-termists, can have a realistic impact on cultural and publication norms. The before-mentioned target group could take these results into account and in a mutual effort try to solve the problem of collective coordination to maximize expected good.

The method of logical proof should be an accessible and suitable method for the technically adept such as the AI engineer and AI researcher. Furthermore, as outlined in the previous section (1.4), the assumptions of our argument include formal components such as Venn-diagrams, multi-dimensional spaces, expected value terminology and probabilistic processes. Engineers and researchers should be used to handle such concepts as it is part of their education and for most, it is their daily bread and butter.

# 2 The Reconciliation Argument

After introducing some short-hand notation, we introduce the full argument and then analyze each step of the argument separately.

## 2.1 Definitions

To keep formulations compact, we introduce some short-hand notations for recurring terms.

**Definition 2.1: NT**

NT are people that strongly prioritize near-term issues in AI like cyber-attacks, autonomous weapons, technological unemployment, algorithmic bias, deepfakes, autonomous vehicles and many more.

**Definition 2.2: LT**

LT are people that strongly prioritize long-term issues in AI like the development of AGI and ASI.

**Definition 2.3: NT-LT-game**

The NT-LT-game is the game-theoretic situation between NT and LT competing for resources to further their corresponding agenda.

## 2.2 The Argument

Keeping in mind these definitions, let us face the main argument of this paper. The Reconciliation Argument can be found in the following box. Let us break it down point by point. This argument has a very simple structure. It contains three assumptions (1, 2 and 3). The first two are implications while the third one is a proposition. 4 and 5 are consequents that are obtained by applying modi ponens.

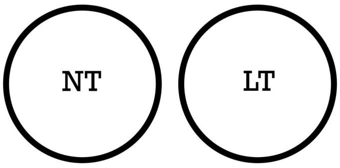
**The Reconciliation Argument**

1. If the NL-LT-game is clearly not a zero-sum game, NT and LT can be reconciled.
2. If both, NT and LT have a significantly higher expected value when not focusing exclusively on their main issues but also taking into account the other side's issue, then the NT-LT-game is clearly not a zero-sum game.
3. Both, NT and LT have a significantly higher expected value when not focusing exclusively on their main issues but also taking into account the other side's issue.
4. From 2 and 3, it follows that the NT-LT-game is clearly not a zero-sum game.
5. From 1 and 4, it follows that NT and LT can be reconciled.

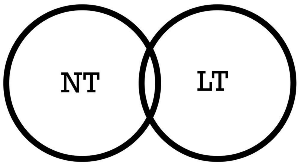
Next, we explore each individual step of the argument.

## 2.3 The Reconciliation Implication

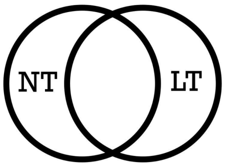
The first assumption seems to be fairly unproblematic. Two parties can only be reconciled if not all actions of one player are detrimental to the interests of another player and vice versa. Thus, if there are actions that do not only further one player's own interest but further (or at least keep the state neutral) for another player, both players can be reconciled. Admittedly, this implication could be exposed to a reductio ad absurdum. An advocatus diaboli could point towards a scenario which is almost entirely a zero-sum game. Imagine a situation in which there are plenty of possible actions but only a single action is not detrimental to the interests of another player. But this action has only a minuscule positive effect for the other player. All other actions are characteristic actions of pure zero-sum games. Given this scenario, it would be hard to concede a reconciliation of the two players. There being one action among many, which is only slightly positive for another player, is not an intuitive guarantee for reconciliation. For this reason, the formulation of the assumption includes "*clearly* not a zero-sum game". Figure 1 shows the differences between three different kind of games. Figure 1a shows a pure zero-sum game situation where the goals of NT and LT are not overlapping. Figure 1b shows a non-zero-sum game. In this case, it is mentionable that there is indeed an overlap of goals between NT and LT, however, the overlap is relatively small. The game that the before-mentioned advocatus diaboli referred to is this game in figure 1b. Figure 1c shows a clearly non-zero-sum game where the overlap between NT and LT goals is non-trivial. This last situation is the one that is needed in the first assumption of the argument and that would counter the concern of the advocatus diaboli.




(a) Pure Zero-Sum Game




(b) Non-Zero-Sum Game




(c) Clearly Non-Zero-Sum Game

Figure 1: Differences between games

### 2.4 The Non-Zero-Sum Game Implication

Such as the first assumption, the second assumption is also an implication. The antecedent describes a conjunctive situation where two conditions have to hold:

1. NT has a specific level of expected value when focusing only on an NT issue and its corresponding near-term impact on society. NT's expected value is significantly higher if NT considers LT issues into their decision.
2. The same has to hold for LT. Its expected value is significantly higher when considering NT issues into their decision.

Irrespective of these two points being actually true at this moment of analysis, their conjunction makes up the antecedent. From it follows the consequent which states that the NT-LT game is clearly not a zero-sum game. This implication seems to comply with a general intuition about zero-sum games. The expected value of both players increases when not working narrowly on their own issues but taking the goals of the other player into account. In contrast, a zero-sum game does not allow for such synergistic effects.

### 2.5 The Expected Value Assumption

The third assumption of the Reconciliation Argument is a very crucial one. It is the assumption that the antecedent of the implication in assumption two is true. Both, NT and LT have a significantly higher expected value when not focusing exclusively on their main issues but also taking into account the other side's issue. The meaning of this assumption has already been explained in the previous section. Whether this assumption is indeed true, cannot be discussed within a few sentences. In order to show that this assumption is true, we have to introduce more arguments. As this assumption is a conjunction, we can develop two arguments, each tackling the two propositions separately. The two arguments can be found in sections 5 and 6, respectively.

### 2.6 The Modi Ponens

Both, step 4 and 5 of the Reconciliation Argument are results of applying modi ponens. Step 4 states that we take the non-zero-sum game implication (2) and the expected value assumption (3) and as a result it follows that the NT-LT-game is clearly not a zero-sum game. Furthermore, we have step 5, where we take the resulting proposition of step 4 and combine with the reconciliation implication (1) to result in the final proposition of NT and LT being reconcilable.

The structure of this argument is a simple one. The formulations of the assumptions and the implications are chosen in such a way that they fit exactly. Thus, applying the modi ponens in this case is not a problem. See the box below that shows a simplified version of the Reconciliation Argument.

**The Simplified Reconciliation Argument**

1. If Y, then Z.
2. If X, then Y.
3. X.
4. From 2 and 3, it follows that Y.
5. From 1 and 4, it follows that Z.

Given this simplified version, it should become clear that the argument is a sound one. We have elaborated on how assumptions 1 and 2 are of reasonable nature. What is left is to show that the expected value assumption (3) is indeed true. For this purpose, we have to formally introduce two main concepts – openness and path dependence.

## 3 The Openness Space

In order to delve into the remaining arguments, we need to go deeper into one specific NT issue which in our case is openness. To understand the arguments around openness, we need first to chart the openness space. In the next subsection, we elaborate first on the various dimensions of openness. In section 3.2, we look at the expected impacts for NT and LT in an isolated way. In other words: We sketch what kind of openness constitutes a high value from a more narrow perspective of NT on the one hand and of LT on the other hand.

### 3.1 The Dimensions of Openness

At first glance, openness might make the impression to be a binary variable. But this view would be far too simplistic. From a more detailed perspective, openness can be represented as a multi-dimensional vector space indicating how open specific aspects of AI development are. These various aspects of AI development can be more or less open. Given the literature on this matter [23]–[25], we collected some relevant aspects from various researchers to shape a new space of openness.

Given a multitude of openness aspects, some categorization would be convenient. Viktoriya Krakovna suggested such a categorization by distinguishing a technical and a strategic category of openness [24]. All here discussed openness aspects can then be neatly assigned to one of these two categories. An overview of all aspects can be found in table 1 below. All aspects are discussed in the subsequent sections.

| Technical Aspects | Strategic Aspects |
|---|---|
| Science | Capabilities |
| Source code | Organizational Goals |
| Data | Forecasts & timelines |
| Safety techniques | Composition of ethics boards |
| Development tools | Timing of release |
| | Distribution of release |

Table 1: Aspects of technical and strategic openness

### 3.1.1 Technical Openness

Nick Bostrom identified a few of these technical openness aspects which he calls source code, science, data and safety techniques [23]. Let us have a closer look at them individually.

#### 3.1.1.1 Science

In the context of this paper, science can be roughly understood as papers or sometimes even blog posts [26]. The usual modus operandi of academic AI researchers leads to publications of their work. These publications often involve theoretical explanations, algorithms[5] and objective functions.[6] Researchers in the industry are often more restricted in publishing their research results. However, it is not uncommon to find research labs of big companies to make their research publicly available via platforms like arXiv [28] (which is a widely known repository of electronic pre-prints), blogs on their websites or in standard journals.

#### 3.1.1.2 Source Code

Source code can refer to various things such as hyper-parameters, the complete code base or trained models. Often, it will be very hard for another researcher or engineer to recreate a functioning program given only the information provided in a research paper [24]. Fine-tuned *hyper-parameters* in specific ML models are rarely included in the papers.[7] Researchers could, however, go even further and decide to make their entire *code base* publicly available by providing it in some specific programming language on a platform such as GitHub [29] who offers hosting for software development. And although one can find plenty of coding projects on such platforms, it is rare that researchers publish their code along with their state-of-the art research papers.[8] Another aspect of source code could also be a *trained model* if applicable. Especially in deep learning and reinforcement learning frame works, it is common to train one's models which can result in huge files that contain up to billions of trained parameters.[9] A researcher or research team could decide whether they want to publish their hyper-parameters, their trained model or their code-base rather independently.

#### 3.1.1.3 Data

Data is very often a key to good models [33]. And while a lot of data sets are publicly available, a research team can in principle create their own data set and decide whether they want to make it partially or fully available to others or not.[10] Furthermore, platforms like Kaggle [34] offer plenty of data sets for all people to train their ML algorithms on. Other organizations like OpenAI provide the research community with training environments like Gym, Roboschool or Universe where one can train one's reinforcement learning algorithm on a variety of problems [35]. These can then be used as benchmarks to compare performances with others. Such benchmarks could be considered as either data or development tools. See the next but one section.

#### 3.1.1.4 Safety Techniques

Furthermore, Nick Bostrom discussed safety techniques that are subject to the question of openness. Instead of focusing on pushing accuracy and performance values of ML algorithms, some focus on developing safety techniques that can be used in a broad range of ML research and applications. Such safety techniques can for example revolve around avoiding negative side effects, avoiding reward hacking, scalable oversight, safe exploration and robustness to distributional shift. For more details on these safety technique sub-fields, see [36]. Furthermore, the or-

[5] Usually, one can find something that resembles pseudo-code rather than proper code in a specific runnable programming language.

[6] Objective functions in AI or Machine Learning (ML) are usually a formal specification of the quality of a solution [27, p. 121]. If an objective function is a good one, it tells you how good a given solution is. This is especially important for optimization problems. Objective functions can often be found as precise mathematical formalizations in papers.

[7] Examples for common hyper-parameters in e.g. deep learning are thresholds, number of hidden units, learning rates and biases. Such hyper-parameters have to be chosen very carefully to ensure a high performance level.

[8] Some journals such as *Nature* put regulations in place that require the authors of papers to make their executable code available to editors and reviewers of the *Nature* journal [30]. This practice, however, does not make the code available to readers of the journal.

[9] See e.g. the currently released version of OpenAI's language model GPT-2 with 1.5 billion parameters [31] or the currently world's biggest model – Nvidia's BERT model that exhibits 8.3 billion parameters [32].

[10] However, there might be limitations given the nature of the data. If the data contains sensitive information and releasing it to other researchers would violate some form of privacy, the researchers' hand are tied.

ganization *Partnership on AI* has recently released an reinforcement environment that is designed to test the safety of reinforcement agents and algorithms [37]. This could be considered as a combination of safety techniques, data and development tools.

#### 3.1.1.5 Development Tools

We introduce a new technical aspect which we call development tools. If researchers would have to start to code from scratch every time they start a new project, progress would be slow. In early times of computing science, this was the indeed the case. But nowadays it is common to use *libraries* that can include classes, data, algorithms and a documentation for easy use. One can easily install libraries and their ready-made functionality without the need to program all kind of sub-routines. While libraries have long simplified the lives of researchers in the computational sciences, popular deep learning libraries like *TensorFlow* (since 2015) and *PyTorch* (since 2016) have revolutionized ML research and engineering.[11] With the help of such libraries, one can quickly and easily develop and implement new algorithms without having to program central aspects like back-propagation or specific neural network architectures (e.g. CNNs and LSTMs). In contrast to many other open-source applications, TensorFlow and PyTorch have been developed by big companies such as Google and Facebook, respectively. These companies decided to share not their ML code itself but their ML development tools with the world. It seems plausible to believe that many other tools (such as further libraries) are likely to still be only distributed within a company.

### 3.1.2 Strategic Openness

The strategic category of openness includes capabilities, organizational goals, forecasts, the composition of ethics boards, the timing of releases and the distribution of releases. The first two aspects are borrowed from Bostrom [23], forecasts are taken from Krakovna [24] and the latter two have been discussed by J. Whittlestone and A. Ovadya [25].

#### 3.1.2.1 Capabilities

The aspect of capabilities represents how open researchers, research labs or entire companies are about the performance state of their ML algorithms. Often such capabilities can be communicated in the results section of their papers. Especially, academic researchers do so. Major tech companies such as Google, Facebook, Microsoft and Baidu make some of their results also public via papers and poster sessions at conferences [23]. Companies can signal their ML capabilities also via the performance of their products.[12] But industry research labs can also be more secretive about the state of their art. The company *X Development* is a good example which has formerly been known as *Google X*. It is a semi-secret research facility working on a plethora of projects. Little research relevant information reaches the public [39]. Other companies do not share what their ML algorithms are able to do or how well they perform because they are in constant competition with other companies. The current landscape in showing capabilities looks rather mixed.

[11] There are of course many other mentionable DL libraries. For an overview, see [38].

[12] See e.g. the level of voice recognition of Apple's Siri or Google Now.

#### 3.1.2.2 Organizational Goals

Next, we have organizational goals. While individual researchers usually do not have officially stated goals, research labs and companies often do. Organizational goals are often synonymous to mission statements or statements of purpose. And while statements of purpose and mission statements are often very high-level formalizations of their goals or sometimes merely the organization's philosophy or values,[13] organizational goals can also be understood in a broader sense. According to a broader understanding, an organization could state the goals of each of their projects and by transition also the existence of their projects.

#### 3.1.2.3 Forecasts

Forecasts can be considered to be connected with capabilities and organizational goals. A researcher, a research team or a company can make predictions on the basis of their own current capabilities, their goals, their organizational infrastructure and talent source. These predictions about their own future capabilities can then be shared with the research community or not. Furthermore, if more such forecasts exist, meta analyses could be conducted to view forecasts across the research community and get a potentially better approximation of what to expect from the field within th next years.

#### 3.1.2.4 Composition of AI Ethics Boards

Furthermore, we can consider the composition of AI ethics boards as a further strategic aspect of openness. While more and more companies have pledged to establish AI ethics boards throughout the last years [42], the composition of these boards is not always clear. Take for example DeepMind's ethics board whose members have remained a secret since its creation in 2014 [43]. Releasing the information of how one's ethics board is composed is subject to how open a company wants to be.

#### 3.1.2.5 Timing of Release

The second last aspect is the timing of release. One could release any of the technical openness aspects at once, not

[13] See for example Google's mission statement: "Our mission is to organize the world's information and make it universally accessible and useful." [40]. While this statement seems to be to the point, it might be too narrow. Google has its specialization in a search engine but also in online advertising technologies, cloud computing, software and hardware. Other company's mission statements can be broader like the one of OpenAI: "OpenAI's mission is to ensure that artificial general intelligence benefits all of humanity." [41]. This statement is rather broad and leaves plenty of room for interpretation. By choosing how long and detailed the mission statement is, an organization can decide how open they want to be about their goals.

at all or in stages. OpenAI's language model GPT-2 is a recent example for a staged release. Given a prompt, GPT-2 is able to generate coherent paragraphs of text [44]. Instead of releasing the code and the trained model immediately, the OpenAI team decided to release their work in stages. They developed four models which differ in size (124, 355, 774 and 1,500 million parameters). Between February and November 2019, they released the models in this order [44]. More detailed reasoning will be discussed later. For now, it should suffice to view the dimension of timing as a non-binary variable as the GPT-2 case demonstrates.

#### 3.1.2.6 Distribution of Release

The final aspect is the distribution of release. A researcher can make a decision with whom he or she wants to share any of the technical openness aspects. Figure 2 shows the various expanding circles. The brighter the color becomes, the more open is increased. In the middle is the researcher who can keep any part of the research just to him- or herself. Given that the researcher is in a research lab, he or she can share aspects of the research with the lab. Given that the researcher is established at a university, the faculty and then the university would be the next layers of the circles. If the researcher is in the industry, the next circle would be the company. The most outer circle is the general public or the remainder of the world. There is another ellipse which stretches over almost all circles. Instead of sharing aspects of one's research with specific populations (other circles), a researcher can also give access to a person or group of people that requests access. That is what we call *on demand* and it could be any person or group of people in any circle.

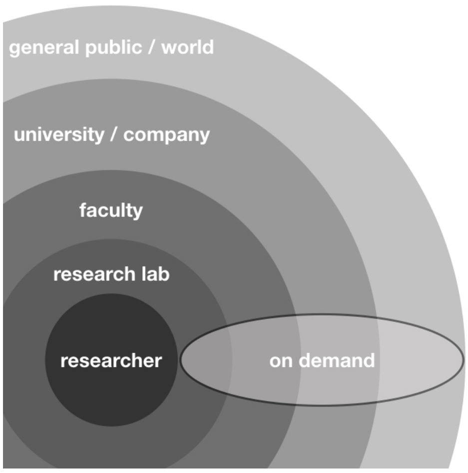



Figure 2: Possible circles of sharing

Researchers in the industry might have a relatively stronger incentive to share less, given the nature of research and development within the company and proprietary software. Academics, however, have a strong incentive to share more. Academic researchers have to apply for grants that finance their research. Certain deliverables such as research results are expected to be shared with the world. Among other considerations, funding is allocated to research projects in such a manner that societal use is increased in expectation. Without sharing the research, there might be no societal use and perhaps no indication that the funding had been allocated wisely.

## 3.2 The Expected Impacts

After having set up an overview over the various dimensions of the openness space, we discuss in the following subsections how traversing interesting parts of this space can have a societal near-term impact and how it can have a societal long-term impact, respectively. Note that at this point, we do not discuss interdependencies between the near-term and the long-term but instead provide an isolated overview of how tweaking specific aspects of openness can exacerbate or ameliorate the near-term or the long-term impact. Only in later sections, we introduce finer nuances and connections between the near-term and the long-term impacts. This section should serve as a provider of an intuition about the two kind of impacts. Note that a huge part of this section is borrowed from Nick Bostrom's analysis [23]. It was not our intention to greatly extend or rectify Bostrom's analysis within this paper. Instead, we use his analysis as a foundation in our later application of path dependence reasoning to openness.

### 3.2.1 Expected Near-Term Impact

Bostrom poses two questions to analyze the near-term impact of openness [23]: Does openness accelerate AI development and deployment and is such an acceleration societally beneficial? We elaborate first on the current state of openness. Then, we consider the positive effects of acceleration, potential negative applications and the first contemporary signs of caution through staged releases.

#### 3.2.1.1 The Current State of Openness

At the current state, Bostrom argues that we have already a relatively high but not maximal degree of technical openness. Most technical aspects such as science, data, safety techniques and development tools are highly open.[14] As mentioned in section 3.1.1, the source code is mostly not openly shared. If the research is shared, however, it usually happens all at once (timing of release)[15] and is shared with the general public (distribution of release). Information about capabilities is generally shared with the public via papers, conferences, blog posts and through demonstrating the performance of ML applications on the market. Forecasts are currently not formally provided by research labs or companies. There are, however, some researchers that try to model an expert-opinion on time-

[14] One could argue that this might apply for science in general but it might be not too big of a stretch to concede the existence of nationally funded, probably military-related research in countries such as Russia, China, Iran and Israel. And although this might be true, we can only speculate about the extent of such research projects. In the context of this paper, we exclude this from our analysis.

[15] A mentionable exception is OpenAI's staged release of their language model GPT-2 over the year of 2019.

lines of general capabilities [45]. In this survey, experts have been asked to provide their probability distributions for the arrival of AGI and its likely impact on humanity. These forecasts are very broad, based on personal intuitions and extend only over the long-term. Clearly stated near-term predictions are more rare. Moreover, some of the big tech companies like Microsoft [42], Facebook [46], Google [47] and DeepMind have some kind of AI ethics board, an ethics committee or some other internal team that is responsible for ethical restrain. However, while the composition of Google's ethics board had been known, the membership of some board members caused unrest and this led eventually to an early cancellation of its AI ethics board. The composition of Facebook's "special team" on ethics is unknown [46]. At the moment, we know that Microsoft's internal committee is lead by Eric Horvitz and his team consists of researchers, engineers and some legal counsel. Furthermore, DeepMind's AI ethics board members have remained a secret since its creation in 2014 [43]. From these cases we can see that the composition of the AI ethic boards and committees is quite closed to maximally closed.[16]

#### 3.2.1.2 Positive Effects of Acceleration

Bostrom considers openness to accelerate the development and deployment of AI in the near-term [23]. According to him, increased openness would lead to a quicker spread of current state-of-the-art techniques which would also lead to a quicker (commercial) application. Sharing technical aspects like science, code, data and development tools are very effortless to share at a marginal cost. And an acceleration of AI development and deployment has probably a net-positive expected impact on society. Economic sectors that would be affected are transportation, healthcare, the environment, entertainment, security and scientific discovery [23]. Take for example (partially) self-driving vehicles: Earth has an annual death toll of around 1.2 million people. Most of them could be avoided in the near future by using more AI in such vehicles [49]. Or take healthcare where AI tools can have an impact on digitization, engagement and diagnostics. Operational processes can be made less expensive. The interaction between patients and consumers on the one hand and healthcare providers, systems and services on the other hand can be improved. And new AI augmented products and services can simplify and enhance diagnoses [50]. A third example can be borrowed from scientific discovery. ML can here be used to gain new insights from observational or even simulated data [51]. By using ML algorithms, researchers in natural sciences can infer causal relationships and predict e.g. properties of chemical compounds and other materials. Common application domains are chemistry, biology, medicine, physics, engineering and material sciences [51].

#### 3.2.1.3 Negative Applications

Generally, more openness seems to be a positive factor for near-term impact. There are however some exceptions such as military uses,[17] applications for social control[18] and systemic risks from more reliance on complex autonomous processes.[19] Given the mostly negative effects of these technological applications, most people would probably not demand the source code for AWS to be made publicly available. Consequently, maximal openness does not seem to be the most desirable path for the near-term. Even if we assume that exclusively military-related AI research is kept in dark by national military research facilities, we are stuck with the problem that AI is a multipurpose technology and foundational research on AI can be used for all kind of applications – including e.g. military applications.

#### 3.2.1.4 First Signs of Caution by Staged Releases

There is one especially interesting event in 2019 that is relevant in the discussion of the near-term impact of openness – OpenAI's staged release of their language model GPT-2. As GPT-2 can create coherent text given a topic, it has positive potential for AI augmented writing systems, dialogue agents, unsupervised language translation and improved speech recognition [52]. However, it has also an unparalleled potential for misuse such as the generation of fake news, online impersonation of others, automated production of text to abuse marginalized groups on social media and the automation of spam and phishing content production [52]. Due to the potential negative consequences of an immediate release of their data set, the training code and the trained model, OpenAI decided to give its researchers more time to conduct more thorough risk-benefit analyses, the general public the possibility to adapt to such new technological progress and other researchers to figure out ways of mitigation of misuse [44]. As time passed by and the risks seem lower than originally expected, OpenAI released more and more powerful versions of their language model.

While there certainly are potential negative AI uses and a strong opacity with respect to secretive state and military research, Bostrom argued that the net-impact for the near-term is rather positive [23]. Nevertheless, first signs of increased caution are emerging trough a precedent of a staged release.

[16]Next to the question about the openness about the composition, it is currently also not transparent how much actual power such AI ethics boards exhibit. We can expect that decisions are handled internally such that the public will usually not know about potential interventions, let alone their details. In this context, we could point to a case at Microsoft who reported to have cut off significant sales on the basis of recommendations given by their internal committee [48]. It remains unclear what the exact matter at hand was or what ethical considerations have led to this result. Due to the lack of openness, it also remains unclear what power such ethics boards have and whether introducing such boards is just a means of ethics washing.

[17]E.g. the usage of autonomous weapon systems (AWS).

[18]E.g. surveillance applications like facial recognition and AI augmented censorship.

[19]E.g. advanced technology enables increased accident risk in economy and infrastructure which could be exploited by parties with malicious intentions.

### 3.2.2 Expected Long-Term Impact

Bostrom presented two challenges for the long-term impact – the *control problem* and the *political problem* [23].

**Definition 3.1: The Control Problem**

The control problem describes the difficulty of designing AI in such a way that it performs as intended by its creators.

**Definition 3.2: The Political Problem**

The political problem describes the challenge to make various actors (like individuals and institutions) further the common good when using AI.

We can now look at the main effects that openness could have on these two problems. Bostrom identified three such effects:

1. Openness accelerates AI development
2. Openness renders an AI race more competitive
3. Openness facilitates wider engagement

Figure 3 below summarizes the impact of each main effect of openness. Red colored arrows represent a negative impact, green arrows represent a positive impact and gray arrows represent a mixed or ambiguous impact. The thickness of the arrows is correlated to the strength of the impact. In the following, we describe these three main effects and how they influence the two problems.

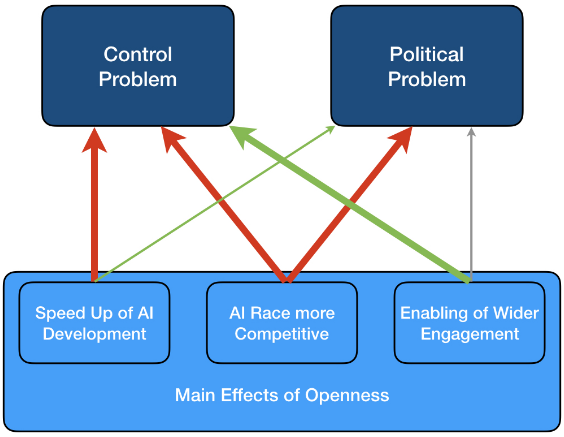



Figure 3: The impact of openness on the control and political problem

#### 3.2.2.1 Openness Accelerates AI Development

According to Bostrom, we can look at the long-term impact in two ways – one where we focus on the currently existing people and one where we also consider the vast amount of people from future generations [23].

In the former case in which we we value currently existing people substantially more, one could argue that more openness would have a positive impact on the long-term. If we assume that more openness does accelerate AI development and deployment, then the emergence of more advanced AI and its benefits will come along sooner.[20] Given that more advanced AI technology (especially ASI) could have a considerable impact on reducing death rates and on ameliorating the wellbeing of potentially all humans, a sooner arrival of ASI would be desirable from a global point of view [53]. Thus, currently existing people would benefit from more openness and by transition from increased acceleration of AI development and deployment.

Let us look at the other case in which we also consider people from future generations. Here, we could say that an acceleration of AI development (which is caused by openness) could lead to a situation in which humanity has less time to prepare for a safe arrival of ASI [23]. Openness of safety measures is already relatively open and would stay so when source code and data would increase in openness. Thence, the relative increase in source code and data openness would give the research community less time to catch up on safety research, everything else being equal. This would impact the control problem in a negative way. The effect on the political problem seems hard to estimate at this point.

Furthermore, there is another positive effect that is associated with earlier ASI arrival – the mitigation of other existential risks, including both anthropogenic and naturogenic existential risks. In summary, the first main effect (openness may accelerate AI development) has a rather weak positive long-term impact.[21] Note again that this is a rather qualitative, not a quantitative assessment.

#### 3.2.2.2 Openness Renders the AI Race More Competitive

This second main effect has a negative impact on the control and the political problem [23]. According to this effect, more openness would probably tie a leading developer's hands when it comes to pausing the development and thus potentially lowering the performance of his or her AI, given a strong competitive landscape where competitors might exploit any advantage that they can get. This would lead to a situation where the control problem is harder to solve. And as a consequence, existential risk would be increased.

The impact on the political problem can be divided into three parts. First, there might be a small decrease of the probability for a situation in which a small group creates a monopoly on the benefits of advanced AI. This would be a small positive impact on the political problem.

Second, in a hardware-constrained scenario,[22] more

[20] Provided that ASI turns out friendly.
[21] Given that we consider future generations into our calculation.
[22] A software-constrained scenario is a situation in which the bottleneck of AI advancement is the development of better software. This scenario is opposed to a hardware-constrained scenario in which access

openness would lead to a situation where advanced AI would be controlled by actors in a proportion that represents currently existing power structures, rather than a single actor (like a company or a research lab). This would be the case because the groups that currently have the money and therefore the hardware,[23] would have an advantage over the the groups that do not have access to hardware resources. If we prefer a wider distribution of powers over a single actor having full control over advanced AI, then we would want more openness. But remember that this would only be the case if we live in a hardware-constrained world. In a software-constrained world, more openness would probably lead to a control monopoly of a single actor as one (or a short series of) software breakthrough(s) could lead to the creation of AGI or ASI. Due to our lack of knowledge about what world we are living in, the impact of this second main effect on the political problem stays ambiguous for now.

Third, a higher competition in the AI race would reduce the probability of a singleton.

**Definition 3.3: Singleton**

A singleton is a world order in which a single decision-maker is at the top of the hierarchy [3, p. 78]. In the context of this paper, we imagine an ASI to be such a top decision-maker.

A reduction of the probability of a singleton scenario is considered to have a negative impact on the political problem [23]. The counterpart to a singleton situation is a multipolar outcome which brings along two main concerns. The first concern implies several superintelligent agents that play against each other. Bostrom hypothesized that offence might have a strategic advantage over defence at some level of technological development [23].[24] In order to reach their mutually non-congruent goals, ASI agents could make use of biotechnology, cyberwarfare, molecular nanotechnology or unprecedented technology that could be invented by such ASI agents [23]. If offence might be favored, the catastrophic and existential risks are substantially increased. Furthermore, the second concern addresses a population explosion of suffering digital minds [54], [55].[25] Without global coordination or regulation, it could be the case that a plethora of digital minds are created and copied to serve as workers that would generate revenue for the creator or owner of these digital minds [56]. Imagine a ruthless employer or even a slave owner that is unbound by any laws, regulations or even conventional human morality. Such a person could squeeze out a lot of profit from his workers with a disregard to their happiness or suffering. An ASI could could do so even more with its digital minds. It seems not entirely unlikely that such digital minds could live miserable lives filled with suffering.[26] One should keep in mind the scope of such suffering as the number of digital copies can be humongous. Copying such minds seems rather inexpensive and quick which is why it could lead to a population explosion and by transition to explosion of suffering. Whether we should estimate such a scenario to be likely is not a simple question. At this moment, we think that it sounds plausible given a lack of global coordination and potential discriminatory sentiments with regard to substrate.[27]

To summarize: openness could make the probability of a multipolar scenario more likely which could lead to an explosion of digital suffering due to a lack of global coordination.

#### 3.2.2.3 Openness Facilitates Wider Engagement

This third main effect has a positive impact on the control problem and a mixed impact on the political problem [23]. Increasing the openness of the aspects of science, data and source code could enable more other researchers to participate working on state-of-the-art systems. One advantage of this scenario could be that such external researchers are less biased or bring in values that correspond more with the public good than the values of e.g. a profit-maximizing tech company. This would improve the situation of the control problem. Another positive factor on the control problem is the enablement of better social planning and prioritization [23]. This would be the case given openness about capabilities, organizational goals and timelines. When the public and the governmental institutions know about the current state of research, they can also infer the timelines for the deployment of products that make use of this technology. This could help to enable better policy adjustments.

The impact of a wider engagement on the political problem is less clear. The openness of capabilities and timelines could also raise the probability that the government will take control over advanced AI [23]. If, however, science, source code and data are opened up more, the probability would decrease that the government takes control because other actors could make use of these resources as well. How to assess the desirability here would depend strongly on counterfactual situations. If the government is not in control, a research team or a company could be instead. If we assume for a moment that it might be better that the government is in control, then the impact stays unclear. It is dependent on whether we make only capabilities and timelines more open or in addition science, source code and data.

and development of hardware is the bottleneck.

[23]Think in terms of computing power, GPU clusters, etc.

[24]It is hard to estimate what the probability for this would be. However, if we assume a non-trivial amount of probability, we have to give these thoughts some weight

[25]A digital mind can have one out of two origins. Either we use whole brain emulation to "scan" an existing human mind or we create a new mind that originates from an AI. We imagine that establishing rights for such digital minds would be harder in the latter case.

[26]Note that this consideration requires the assumption of the digital minds exhibiting sentience. Alternatively, it could also be the case that we have whole brain emulations that approximately can think like humans but not exhibit sentience and therefore do not have the capacity to suffer.

[27]Discrimination with regard to substrate refers to discriminating on the basis of how somebody is implemented – e.g. on silicon or biological tissue.

There is one more interesting point that could affect the political problem. If we increase the openness about capabilities and timelines to the general public,[28] we trigger a wider engagement and a healthy debate throughout our entire society. This could be interpreted positively as all (or at least more) stakeholders' interests might be represented when it comes to deciding how to steer the development of advanced AI. However, an advocatus diaboli could compare this potential situation of including the general public with the debate that has been happening around genetically modified organism (GMO) usage in Europe [23]. One could say that the debate would have been more fruitful if scientists and politicians would have been bigger debating participants than ill-informed citizens. The technological progress in this domain could have been substantially higher without irrationally motivated interference.[29] A similar scenario could be imagined for the debate on (advanced) AI. An ill-informed public could fuel the debate with emotionally charged and irrational arguments which could exacerbate the efforts to find an effective solution to the political problem. Nota bene: We do not necessarily agree that the public AI debate would resemble the GMO debate in Europe. However, the possibility remains. And while the engagement of the public could be positive, there is a non-trivial chance that it could be indeed negative. Thus, the impact of a wider engagement on the political problem remains ambiguous.

In a nutshell, while there are some positive effects on the long-term (especially openness of safety measures, composition of ethics boards and organizational goals), some mixed effects (through open capabilities, timelines and distribution), there also some negative effects (through open source code, science, data and development tools). The general impacts can be viewed in figure 3. Overall, general openness is significantly more negative for the long-term than for the near-term.

# 4 Path Dependence

Path dependence has several interpretations. One common interpretation is a broad one which is characterized by the slogan "History matters!" and which describes how prior and current decisions influence the future [58, p.981]. This almost metaphorical account of path dependence is trivial as it is evidently obvious that all events have causes and alternative decisions would have brought along contingent events. Given this widespread but trivial definition [59], [60], there is not much insight to gain. However, there is a more narrow interpretation of path dependence that involves sup-optimal late scenarios in which it can be very costly to change paths that would lead to more optimal scenarios [61]. This interpretation of path dependence exhibits more potential for useful insights as it can also involve formal models.

[28] I.e. we maximize the scope of distribution of these aspects.

[29] GMO rejection and fear came usually along with a institutional distrust (government and companies) and anti-science attitudes. Rational arguments tend to fail the GMO debate [57]

In the following subsection, we introduce some standard examples from the literature to provide a better intuitive grasp of this section's main concept. Next, we analyze the theory of formal models where we distinguish different kinds of path dependence. In section 4.3, we look closer at what kind of path dependent process the traversing of the openness space could be. Finally, we elaborate in section 4.4 why it is probably hard to escape a lock-in situation in a late stage of the process.

## 4.1 Examples of Path Dependence

The literature provides a few common examples for path dependent processes like the QWERTY keyboard layout [22], [61], [62], the videotape format war [62], [63] and the permanent seats on the UN security council [64].

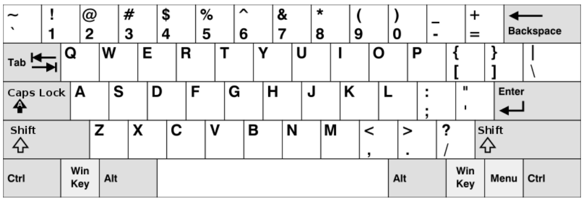
### 4.1.1 The QWERTY Keyboard Layout

The QWERTY keyboard layout (see figure 4a) had originally been designed in the 1870s to be used for type writers [62]. The keys were attached to metal arms which tended to jam if neighboring keys were pressed simultaneously or in quick succession. It is said that the layout was mostly a result from avoiding jams by placing common letter combinations far apart [62].[30] Let us fast-forward 150 years and type writers are practically not used and replaced by modern computers and physical or even virtual keyboards. However, we are still using the QWERTY keyboard layout although jamming of the input machine is not an issue anymore. Arguably, there are potential better keyboard layouts like the Dvorak layout (see figure 4b). However, most computers come with the QWERTY keyboard. Acquiring a Dvorak keyboard necessitates extra effort. So does learning to use a new keyboard layout. These are resources that the vast majority of people is not willing to invest. Thus, we are all stuck with a seemingly sub-optimal keyboard layout. If the first successful type writers would have used the Dvorak keyboard layout, it would have been likely that we would still be using that layout instead of any other layout.

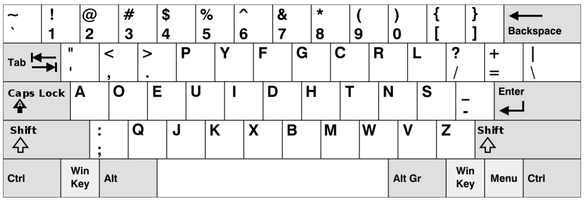
### 4.1.2 The Videotape Format War

The videotape format war was a dominance battle between the VHS and Betamax tapes [62]. And although the quality of the Betamax product was higher than the one of the VHS product, customers opted rather for VHS as the quality difference was negligible to the customers but the price difference was not. Thus, consumers rented or bought initially more VHS tapes which caused stores to offer more VHS tapes and which caused more customers to acquire VHS tapes, etc. The bandwagon effect was at play. At some point, the VHS product exhibited such market dominance that Betamax had no chance anymore and disap-

[30] The actual story is a bit more intricate but this simplification should suffice for our analysis.




(a) QWERTY keyboard layout [65]




(b) Dvorak keyboard layout [66]

Figure 4: Keyboard layouts

peared entirely from the market. If the price difference would have been insignificant or if Betamax would have spread a bit quicker in the initial state, it is possible that it would have acquired market dominance instead [63]. However, minor initial conditions have lead to a path from which there was no coming back from.

#### 4.1.3 Permanent Seats on the UN Security Council

The third example that we would like to present is the distribution of permanent seats on the UN security council. With its creation, the UN established the UN security council in 1945 and granted the allies in World War II permanent seats on this council. These countries include China, France, Russia, the UK and the USA. This has not been subject to any change since its creation. Although this assignment of permanent seats might be considered as a sensible decision for the immediate post-war era, many would argue that it is not an ideal distribution for the current era. Changes that include the replacement of some members with others or to establish additional permanent seats have been suggested. But consensus has not been found yet as some countries oppose potential assignments of permanent seats to other countries. And although most agree that the status quo is definitely sub-optimal, the situation seems to be locked-in [64].

### 4.2 Theory of Path Dependence

Scott Page provides a formal account of path dependence in his 2006 paper [22]. We will be mainly using his elaborations to provide an overview of path dependence and its aspects that are relevant for this paper. For this purpose, we introduce various dynamic processes that have some flavor of path dependence which we will formally define.

#### 4.2.1 An Overview of Dynamic Processes

Table 2 provides an overview of some of the various models which Page introduced [22].[31] We explain these models in isolation. Their complexity increases as they are usually building upon their predecessors. Later we check which of these models could represent the openness process[32] the best. Do not worry if not all properties are self-explanatory. We discuss all of them in great detail in the subsequent sections.

| Process | Properties |
|---|---|
| Bernoulli | Independent |
| Polya | Phat dep., multiple equilibria |
| Balancing | Phat dep., unique equilibria |
| Biased Polya | ↑ returns, unique equilibria |
| Strong Path Dependent | Path dependent |
| Founder | Initial outcome dependent |
| Cascade | Early path dependent |
| Unstable Government | Last outcome dependent |
| Forgetting | Recent path dependent |

Table 2: Ball and urn models

#### 4.2.2 Notation and Basic Definitions

Let us start first with some notation. Time $t$ is discrete and indexed by positive integers. Outcomes at some time period are denoted as $x_t$. The history at time $T$ is represented by $h_T$ which refers to the combination of all outcomes $x$ up to and including the time period $T$. The outcome function $G : h_t \rightarrow x_{t+1}$ or $x_{t+1} = G(h_t)$ can be probabilistic.[33]

It is crucial to distinguish *outcome dependent* from *equilibrium dependent* processes as outcomes and equilibria are often wrongfully conflated.[34]

**Definition 4.1: Outcome Dependence**

We say that a process is outcome dependent if the outcome $x_t$ at time $t$ depends on at least another outcome $x_s$ where $s \leq t$.

**Definition 4.2: Equilibrium Dependence**

For a process to be equilibrium dependent, it has to be the case that not only a specific outcome but the long-run distribution over outcomes has to depend on past outcomes.

Page provided an example to distinguish these two dependences [22]. The act of settling the North-American continent in the 19th century is outcome dependent but not

[31] For a detailed account of all processes, see the section *Dynamic Processes* in his paper [22] For our purposes, it will suffice to cover a subset of them which is nine out of eleven.

[32] The *openness process* is from now on the short hand term for traversing the openness space

[33] Where necessary, more notation will be introduced throughout the upcoming sections.

[34] Note however, outcome dependence is a necessary condition for equilibrium dependence

equilibrium dependent. We could imagine an outcome to refer to a map of this continent indicating for each region whether it has been settled already or not.[35] Time $t$ can be represented by years. There are different paths, i.e. where to go next to settle some land. An outcome $x_{1852}$ depends on its history $h_{1852}$, i.e. outcomes $x_{1851}$, $x_{1850}$, etc. Settlers spread throughout the country in a continuous way from one patch of land to an adjacent one. Thus, whether a specific area is going to be settled in the next year, depends simply whether adjacent areas are already settled. The long-run equilibrium however is not dependent on the intermediate steps of settling the continent. We know that the settlement of the continent will eventually be complete (all green); however, it is not dependent on any particular outcome $x_t$, i.e. it is not relevant what areas have been settled when.

To enable a better understanding, we make use of an urn model that contains balls of different colors. At each time period $t$, a ball is randomly selected from the urn. Thence, the outcome function is constituted by the selection of a ball. The outcome $x$ is the color of the selected ball. Given the nature of a particular process, we might add or remove the selected or other balls. The probability of what color a selected ball will have, depends on the composition of the urn. In order to keep the model simple, let us assume that balls can have only one out of two colors – red or blue. Given the notation, the basic definitions and this model scaffolding, we will continue with various processes and their different properties.

### 4.2.3 Bernoulli Process

The most simple dynamic process is the Bernoulli process. At time period $t$, a red ball is selected. Then, the ball is placed back into the urn which leaves the urn's composition unaltered. Therefore, the probability of selecting a red (or blue) ball at time period $x_{t+1}$ is the same as before, i.e. $P_{x_{t+1}} = P_{x_t}$ for all $t$. This process is thus outcome independent.

**Definition 4.3: Outcome Independence**

An outcome $x_t$ is **independent** if $x_t$ does not depend on its history $h_t$ or any other outcome $x_s$ where $s \leq t$.

Consequently, this dynamic process can be formulated as $x_{t+1} = G(\cdot)$.[36] A classical everyday example is a coin-flip.

### 4.2.4 Polya Process

The second process is the well-known Polya process. Imagine that we start with a red and blue ball in the urn. We select a red ball and return this ball to the urn. Additionally, we put one more red ball into the urn. $P_{red}$ increases from $\frac{1}{2}$ to $\frac{2}{3}$ in the next time period. In the subsequent selections, it becomes more and more likely that the urn's composition will be dominated by red balls. The Polya process is a good example for increasing returns which is often conflated with path dependence itself. Increasing returns refer to the phenomenon where selecting a specific outcome (here a red ball) will make it more likely in the future to be selected again [67]. This process is equilibrium dependent. There will be a long-run equilibrium with any ratio;[37] however, this equilibrium is not dependent on the order in which the balls have been selected. Only the set of selected balls is relevant. This makes the Polya process phat dependent.[38]

**Definition 4.4: Phat Dependence**

A phat-dependent process has an outcome $x_t$ that is dependent on its history $h_t$ but is independent of the order of the history's outcomes.

Given this definition, the Polya process can thus be formulated as $x_{t+1} = G(\{h_t\})$.

By contrast, we define path dependent processes as follows:

**Definition 4.5: Path Dependence**

A path dependent process is given by an outcome $x_t$ for any $t$ depending on its history $h_t$ where the order of the outcomes is important. Such a process can then be described as $x_{t+1} = G(h_t)$.

### 4.2.5 Balancing Process

The third process is the balancing process and resembles the Polya process. However, instead of adding a ball of the same color, we add a ball of the opposite color to the urn. If we select a red ball, it will be returned and accompanied by an additional blue ball. Let us say that we start with two red balls and a blue one. Given that $P_{red_t} = \frac{2}{3}$, the probability to select a red ball is higher, we probably select a red ball, i.e. $x_t = red$. In this case, we have to add a blue ball to the urn which changes the probability of selecting a red ball in the next time period, i.e. $P_{red_{t+1}} = \frac{1}{2}$. Instead of introducing increasing returns, we balance out the probability distribution over selecting balls. This process is phat-dependent with a unique equilibrium at a ratio of one-to-one [22]. As an example for the balancing process, consider the selection of locations for big sports events such as a world cup or the Olympics [22].

### 4.2.6 Biased Polya Process

The fifth process is the biased Polya process which will be central for the openness process. We start again with one

[35] For simplicity's sake, imagine a color-code like green and red. A green patch on the map represents that this area has been settled. A red patch represents that it has not been settled yet.

[36] The dot in the brackets represents the lack of dependence on any outcome or history.

[37] For reasons why there are multiple equilibria with *any* possible ratio, see e.g. [67]

[38] There is no typo in the word *phat*. It is used to distinguish it from the word *path*. As will be explained in definitions 4.4 and 4.5, in a path dependent process, the order of the outcomes is important while we do not care about the order in a phat dependent process.

red and one blue ball. Let us assume a bias towards red balls. If a blue ball is selected, we return it to the urn and add a blue and a red ball. If we select a red ball, we return it to the urn and add $2t$ red balls. It is easy to see how red balls have increasing returns. Furthermore, if we continue selecting balls long enough, i.e. $t \to \infty$, the probability of selecting a red ball at a given time period $t$ will converge to 1. Thus, we have a unique equilibrium at hand where the composition of the urn is greatly dominated by red balls.

#### 4.2.7 Strong Path Dependent Process

The next process in table 2 is the strong path dependent process. Like in the Polya process, we start with one red and one blue ball in the urn. At time $t = 1$, we select a red ball. Now, we put the red ball back and add further $2^{t-1}$ balls, i.e. in this case one red ball. Because the urn's number of balls of a specific color increases at such a specific rate, we can reconstruct the exact order of outcomes given the color of the initial balls and the color of all added balls. The sequence of outcomes would be a unique one. This process is strongly path dependent.

**Definition 4.6: Strong Path Dependence**

A strong path dependent process has the property that the outcome function $G$ differs for any two non-identical histories $h_i$ and $\hat{h}_i$ where $h_i \neq \hat{h}_i$ for some $i = 1...t$. Thus, $G_t(h_t) \neq G_t(\hat{h}_t)$.

#### 4.2.8 Founder Process

Next, we have the founder process. This is a very simplistic model with only two paths. We start out with one red and one blue ball again. When we select a red ball, we return it and remove the blue ball. At the next time period, we will be guaranteed to select the red ball again and we have to remove a blue ball which we cannot do anymore. Thus, we are done as every other time period will not change the composition of the urn anymore. This process is initial outcome dependent.

**Definition 4.7: Initial Outcome Dependence**

This is the case if all following outcomes $x_t$ with $t > 1$ depend only on $x_1$, such that the process can be described as $x_{t+1} = G(h_2)$.

#### 4.2.9 Cascade

The next process is called a cascade. We start again with one red and one blue ball in the urn. We select a ball and return it back to the urn. If we select a ball with the same color (e.g. red) in three consecutive time periods, we remove the ball of the other color (blue). This is what we call early path dependence.

**Definition 4.8: Early Path Dependence**

An outcome $x_t$ depends only on the history up to some $T$ with $T < t$. The process can then be described as $x_{t+1} = G_t(h_T)$ for $T < t$.

It is similar to the founder process, however we do not make the path only dependent on one single outcome but on a series of outcomes.

#### 4.2.10 Unstable Government

The founder process has a counterpart which is the unstable government process. Instead of an outcome $x_t$ being dependent on the initial outcome $x_1$, the unstable government process has the property of an outcome $x_t$ being dependent solely on its prior outcome $x_{t-1}$.

#### 4.2.11 Forgetting

Such as the founder process has its counterpart, so does the cascade process. We call it a forgetting process. We start with one red and one blue ball in the urn. If we select a red ball at time period $t$, we add another red ball for $k$ periods where $k$ is some predefined positive natural number. Let us say for example that $k = 3$. Thus, we add another red ball to the urn after selecting a red ball at time period $t$. At time period $t + 1$, there will be two red balls and one blue ball in the urn. The red ball, which we added previously, will remain in the urn up to time period $t + 4$. After that we will have to remove it again. In the meantime, the rule applies also in other time steps. That means if we select a blue ball at time period $t+1$, we add another blue ball to the mix for exactly three more time steps. Like this, an outcome is only dependent on the recent past.

**Definition 4.9: Recent Path Dependence**

An outcome $x_t$ depends only on the recent history of $x_t$. The process can then be described as $x_{t+1} = G_t(h_t/h_s)$ for $s < t$. We exclude all outcomes that came before a specific time period $s$.

The relationship between $s$ and $k$ is thus defined as $k = t - s$.

Having discussed the relevant dynamic models, we will come back to them when we model the openness process as a dynamic model.

### 4.3 Application of Path Dependence Reasoning to Openness

Has openness some property of path dependence? To answer this question, we attempt to apply the dynamic model on openness. Now, the obvious question is: What kind of process is the openness process? Let us first formulate openness in terms of our earlier introduced notation.

#### 4.3.1 General Notation

Although time in real life is continuous and decisions about various openness aspects could be made at any time, we make the following assumption to simplify the model: time is discrete and indexed by years, i.e.

$$t \in \{n \mid n \in \mathbb{N} \wedge n \geq 2020\}$$

This implies that we treat decisions like they would be made once a year. Given a long, potentially decades or centuries long path to ASI, an indexation of time by year seems reasonable. An outcome at time period $t$ is denoted as $x_t$ and represents one configuration of the openness space. Looking back at table 1, we know that we have eleven openness dimensions. For simplicity's sake, let us assume that each dimension $d$ has a value $V_d$ with $V_d \in \{\textit{very low}, \textit{low}, \textit{medium}, \textit{high}, \textit{very high}\}$. An outcome is equivalent to a configuration which is an eleven dimensional vector containing values from this domain. To offer some intuition, table 3 provides two possible outcomes, $x_{2020}$ and $x_{2030}$.[39]

| Dimension | $\mathbf{x_{2020}}$ | $\mathbf{x_{2030}}$ |
|---|---|---|
| Science | high | high |
| Source Code | very low | medium |
| Data | low | low |
| Safety Techniques | high | very high |
| Development Tools | high | very high |
| Capabilities | high | medium |
| Organizational Goals | medium | high |
| Forecasts & Timelines | very low | low |
| Composition of Ethic Boards | very low | medium |
| Timing of Release | high | high |
| Distribution of Release | high | very high |

Table 3: Example configurations of the openness space for the years 2020 and 2030

The history $h_T$ represents the sequence of all outcomes $x_t$ where $t < T$. The outcome function $G$ is defined by the selection of outcomes from a metaphorical urn. Which exact outcome function we have at hand, has to be concluded in the following analysis.

#### 4.3.2 Check for Independence

Let us first check whether the openness process is an outcome independent process.[40] According to definition 4.3, an outcome $x_t$ is independent if $x_t$ does not depend on any prior outcome $x_s$ with $s \leq t$. Intuitively, outcome independence would hold in the openness process if we could theoretically jump from one arbitrary configuration to any other arbitrary configuration. Hence, if the openness process would be truly outcome independent, we could select an outcome with a configuration with values of only *very low* at time period $t$ and then select another configuration at time period $t+1$ which has only values of *very high*. This jump from a very low overall openness in one year to a very high overall openness in the next year seems very unlikely from an intuitive point of view. Consequently, we conclude that the openness process is not outcome independent.

#### 4.3.3 The Transition Rule of the First Step

In the urn model from section 4.2.1, we had only red and blue balls and thus only 2 different outcomes. The openness space has however 11 dimensions with a domain of 5 values per dimension, i.e. the number of different possible outcomes is now $11^5 = 161051$. In the red-blue model, it was the case that after selecting a ball, the rules indicated to add or remove balls of the same or opposite color. Thus, the probabilities $P_{red}$ and $P_{blue}$ were changed. For our openness model, we cannot explicitly state all probability changes for each of the $11^5$ outcomes. Thus, we have to find a different way to adjust probabilities and we have to figure out the rules according to which the probabilities of selecting outcomes change. As mentioned earlier, we can not copy the rules from the red-blue urn model as the number of different outcomes in our openness model exceeds it by a lot. Intuitively, it makes sense to see configurations undergoing small deviations or mutations from one year to another. This is due to the slow process of changing norms in general. Note that a push for more openness in e.g. source code from *very low* to *low* has to be caused by not only one researcher but by a majority of researchers. Changes usually happen gradually and probably over the duration of many publications. A jump from one configuration to another distant one is possible but less likely. Figure 5 captures this change in transition probabilities graphically. The configurations in figure 5 are color-coded. The darker the color of a configuration, the higher is the probability of selecting this configuration in the subsequent time period.

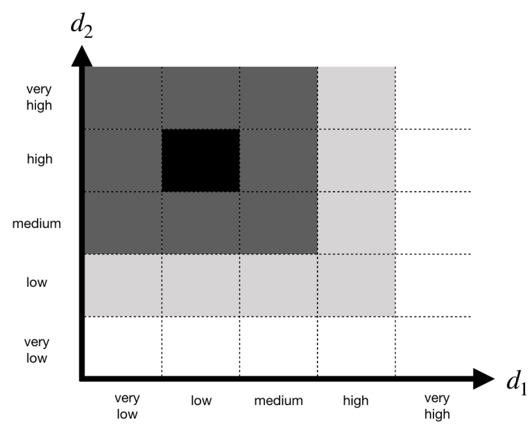



Figure 5: Graphical representation of transition probabilities of openness configurations

Although we can not provide a graphical representation of an eleven dimensional vector space, a two dimensional one should do the trick of conveying the general

[39] The values of outcome $x_{2020}$ are estimated according to our analysis in section 3.1 and the values of outcome $x_{2030}$ are one configuration out of many possible ones.

[40] If this would be the case, the remainder of the analysis would be short.

principle. Figure 5 provides a simplified space with only two dimensions $d_1$ and $d_2$.[41] Both dimensions have values ranging from *very low* to *very high*. The current configuration $x_t$ is represented by the vector $V_{1,2} = (low, high)$. Intuition suggests that adjacent configurations should have a relatively high probability of being selected in time period $t + 1$. Thus, to illustrate an example: Given this new transition rule, the relationship of the following probabilities hold:

$$P_{t+1}(very\,low, very\,high) = P_{t+1}(medium, medium)$$

$$P_{t+1}(medium, medium) > P_{t+1}(medium, low)$$

This simplified two-dimensional space is manageable and one can move rather swiftly (i.e. in few time periods) from one extreme configuration to another. When we imagine the full openness space with its eleven dimensions, it would take statistically many more transitions to move from one extreme configuration to another one. The principle remains however the same – the transition probability mass is centered around the current configuration and decreasing with greater distance in space.

This version of our dynamic openness model is reminiscent of the unstable government process. The current configuration $x_t$ is only dependent on $x_{t-1}$. But this is just a rude simplification and we need to add more nuances to the openness model. We can consider this adjustment rule for probabilities as a good rule for the initial step.

#### 4.3.4 The Transition Rule of Subsequent Steps

Intuitively, an outcome $x_t$ should be dependent on more than its direct predecessor outcome $x_{t-1}$. Assume the following time period ordering: $r << s << t$. Imagine that at some time period $r$, we would select a configuration $x_r$ which has a *data* value and a *source code* value of *medium*.[42] Let us assume that this configuration has been experienced as deeply unsatisfying for society or at least for the AI research community. Next, some considerable amount of time passes by and we reach time period $s$. The configuration $x_s$ is one that has the values *very low* and *low* for the dimensions *data* and *source code*. Some more time passes by and we select another outcome $x_t$. What does our intuition tell us about the likelihood of having steered back into exactly the same configuration as $x_r$ which has been widely unpopular? Despite some years having been passed by, we assume that the memories of "bad" subspace configurations would make them less likely to be selected again. Thus, outcome $x_t$ is dependent on more than its direct predecessor.

Let us assume for a moment that an outcome $x_t$ is uniformly dependent on each outcome in its history $h_t$. We could model this in the following way: We start with some probability distribution over all outcome configurations. Analogously to the probability distribution over the pixels in figure 5, the probability mass is distributed over all hypervoxels[43] in the eleven dimensional openness space. When we select a new outcome $x_{t+1}$, we have to add the new probability mass that this outcome is providing[44] and normalize the new transition values, such that we continue to have probabilities at hand. Given our uniform dependence assumption at the beginning of this paragraph, we would normalize the transition values over all newly added probability masses uniformly.

Intuitively, openness seems to make the impression that the more recent history is more relevant than the less recent one. Being in one high-dimensional "corner" of the openness space seems to be substantially more important to affect the selection of the next outcome than some ancient configuration many time periods ago. Thus, we could introduce a discounting factor $\gamma$ with $0 \leq \gamma \leq 1$. Normalization could be handled by multiplying the factor $\gamma^{t-n}$ with the added probability mass by an outcome $x_n$ where $n \leq t$. In this manner, we could account for a decaying influence of the past. If $\gamma$ would be close to zero, the past would become vastly irrelevant. If $\gamma$ would approximate one, the influence of past outcomes would become more and more equally important. Note that this does not imply that an outcome $x_t$ is *merely* dependent only on its recent history but potentially on its entire history in a discounting fashion.

#### 4.3.5 A Bias Towards Openness

In this section, we attempt to show that the openness process has properties of a biased Polya process. To do so, we have to show that moving into one direction within the space is "easier" than into another direction.[45] Thus, we have to ask: Is it easier to move from an open configuration to a more closed one? Is it the other way around? Or is it equally easy/hard to move into any direction of the space? We identified three reasons why moving from closed to more open is easier than vice versa.

First, there seems to be a general sentiment among software developers and researchers that prefer openness in their modus operandi. Already the early principles of conducting science have had sharing scientific knowledge and its practices at its heart [68]. This principle, although maybe not ideally realized, is nevertheless strongly upheld by a majority of the scientific community. Furthermore, in the early years of computers in the 1950s and 1960s, software was developed and shared especially at universities. The scientific principle of sharing and openness seems to have spilled over into software development. Throughout the decades, code has been formally and informally

[41] $d_1$ and $d_2$ are placeholders for any dimensions of the openness space.

[42] A medium level of openness in data could refer to publicly sharing data, however only partially, e.g. providing a fraction of all records or omitting some dimensions. A medium level of openness in source code could refer to not sharing the entire code base but the values of all the AI-model relevant hyper-parameters.

[43] A hypervoxel is the higher-dimensional equivalent of a pixel or a voxel.

[44] I.e. highest probability mass is in the selected hypervoxel (the configuration $x_{t+1}$) and the probability mass is ever-decreasing for more distant hypervoxels.

[45] And by "easier" we mean a probability distribution favors hypervoxels into a specific direction more than hypervoxels that are on the path of a different direction.

shared. A big open-source culture has developed and resulted among other projects in viable operating systems such as distributions of Linux. Next to altruistic reasons, there is also skill-signalling by especially young, not yet established programmers which plays a role in preferring increased openness [23].

The second reason to consider a bias builds to some degree upon the first reason. We have a historical case where an organization decided publicly and vocally to limit some of its openness aspects. We are referring again to OpenAI's GPT-2 model. The decision to not release their source code and trained model (all at once) earned them quite some backlash from the AI research community and the open source community in particular [69]. One could say that OpenAI's case was specifically interesting as they have the label "open" in their name. The decision to not publish everything at once due to concerns for societal ramifications, has been met with waves of disapproval. Are similar reactions to expect when it comes to other organizations like AI labs in the industry or academic AI labs switching from a currently rather open culture to a more closed one? It would not be implausible to expect similar reactions. This incident could function as evidence that moving from openness to more closedness is connected to the possibility of public backlash. In contrast, an organization that moves from closedness towards more openness, is likely to receive acknowledgement or praise by the research community as it is in accordance with the before-mentioned scientific principle of sharing knowledge.

Third, as soon as information is made openly available once, it is hard to impossible to retract that information. Imagine for instance that some research team has a big breakthrough in some AI application. The team publishes their paper on arXiv and releases their source code and data online on some platform like GitHub. One day later, they realize that this might have been not the best idea. They can delete their paper from arXiv and their corresponding GitHub repository. However, it is likely that the paper and the code have been downloaded various hundreds of times already. If the downloaders realize that the information has been deleted, they might make it available again as a service to the research community. Even if they would know that the original research team wishes their information not to be re-released, it requires only a single unilateral decision by at least one of the downloaders to disregard these wishes and share the knowledge with the world. Similar scenarios could be imagined for other dimensions of openness. Increasing openness is likely to be welcomed while decreasing openness is likely to be shunned. Retracting information that has gone public seems very hard to impossible. Consequently, moving from openness to closedness is hard.[46]

Given these three reasons, we conclude that it is likely that there is a bias favoring more openness. Without looking at a specific configuration, we suspect a gradient in probability transitions as depicted in figure 6. The blue arrows represent the direction of the most probable transitions from any given hypervoxel when no other context (e.g. $h_t$) is known. We assume that it is still possible to move against the direction of the blue arrows, but it is less likely than staying in one hypervoxel or moving along the direction of the blue arrows. If we translate this insight into probability adjustments, we can incorporate them into our dynamic openness model.

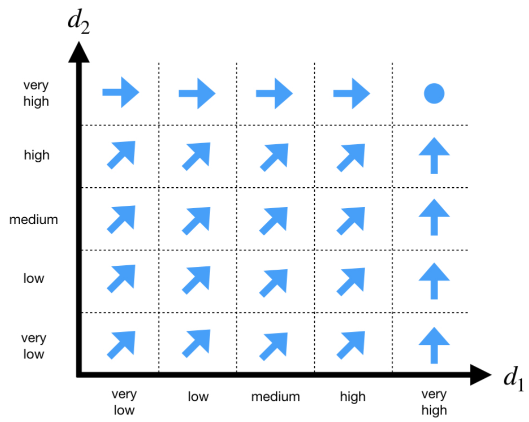



Figure 6: A gradient towards openness

As announced in the beginning of this section, we compare now this version of the openness process with the biased Polya process which had been introduced in section 4.2.1. Given a scaled up gradient that would resemble the one that is depicted in figure 6, we can foresee how over time, i.e. $t \to \infty$, more open configurations would exhibit increasing returns. Eventually, we will converge in a unique equilibrium at a maximally open configuration. The stronger the influence of the bias, the sooner the convergence will happen. If our considerations are correct, the openness process would be comparable to a biased Polya process.

#### 4.3.6 A Summary of the Dynamic Openness Model

In summary, the openness process could be considered to be a biased Polya process and thus an equilibrium dependent process. We first start the selection process with a configuration $x_t$ and a corresponding probability distribution over all outcomes while more distant configurations are less likely to be reached in the subsequent time period $t+1$. Second, not only the last outcome $x_{t-1}$ affects the probability of outcome $x_t$, but its entire history $h_t$ does so while older outcomes are more discounted than recent outcomes. Third, we add a bias that generally favors openness over closedness.

[46] A potential alternative way of censorship could be *flooding* which can be defined as a coordinated attempt of producing and spreading information with the goal of distracting people from accessing the relevant information [70, p. 80]. This is specifically executed by modern regimes like China as a very subtle way of censorship. Similarly, *irrelevant* AI research could be produced to flood the research community. However, this seems harder to execute as it would take a considerable amount of effort to produce a paper – compared to e.g. an online news article. Furthermore, there are journals that ideally consider only relevant research into their publication.

#### 4.3.7 Limitations of the Dynamic Openness Model

Even if the openness process is indeed a biased Polya process, time has to approximate infinity, i.e. $t \to \infty$, such that an equilibrium of maximal openness is reached. If ASI's arrival is likely to be expected within the 21st century [45], not enough time will have passed to guarantee an equilibrium. In this regard, the biased Polya process might be only of limited use for our considerations on action recommendations. Further elaborations on this matter can be found in section 7.

Although we have provided three reasons for introducing a bias towards openness, this bias could potentially be minuscule in reality. It could also be the case that this assumption is just wrong. Maybe only some dimensions are biased towards more openness while others are not. If this entire assumption would be incorrect, we would have potentially a very different dynamic process with different properties at hand as the equilibrium dependence property might disappear. Moreover, our assumption of the discounted history influence could be erroneous. It could be the case that only the recent past counts which would make the process a forgetting process. We have to emphasize at this point that we do not assign a high certainty to most of our assumptions. More thorough scrutinizing will be required.

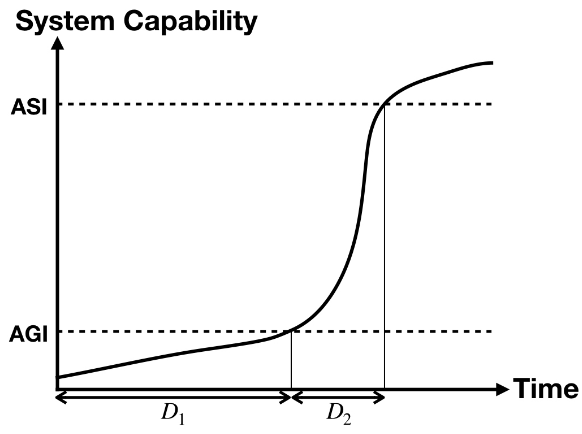
This model is a first attempt in formalizing the openness process and it is very likely that some crucial considerations are missing because we have failed to come up with them. Such missing crucial considerations could change the nature of the model and introduce other properties. Some of these factors could adjust the transition rules or afflict some kind of shock to the probability distributions. It is probable that such considerations can be found if more time is spent on modelling such that more nuances are added to the model.

## 4.4 A Hard Escape from Openness Lock-ins

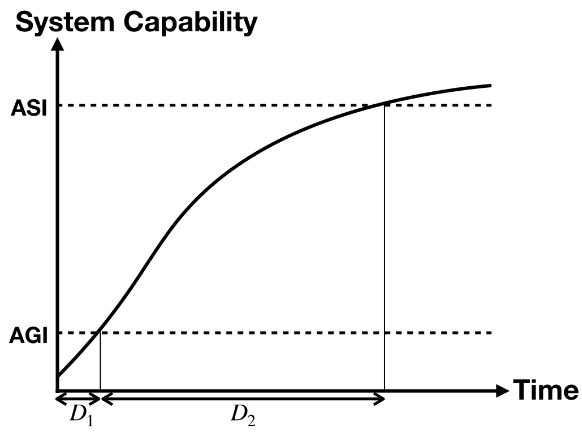
In this section, we elaborate what reasons there are to believe that it is hard to escape an openness lock-in. We start with section 4.4.1 by looking into various take-off durations as the exact kind of take-off is relevant for the potential impact of ASI. In the subsequent subsection 4.4.2, we describe a particular scenario in the future which will help to think through possible ramifications. In section 4.4.3, we describe what reactions could be expected in the introduced scenario and how our model could be adapted. In section 4.4.4, we describe the possible consequences for society. Finally in section 4.4.5, we show why our results should be interpreted with caution.

#### 4.4.1 Take-off Timelines

We have to distinguish two different durations – the *time until take-off* and the *take-off duration*. Figure 7a provides a graphical representation[47] in which the difference should become clear. We plot system capability against time. System capability refers here to the capability of the strongest AI system that humanity has at a given time. We distinguish in particular the levels of AGI and ASI. The time axis starts at the current time. $D_1$ is the time until take-off, i.e. the time from now up to the point where we reach AGI. $D_2$ is the take-off duration, meaning the time from the arrival of AGI up to the point when we reach ASI capabilities.




(a) Hard take-off scenario




(b) Soft take-off scenario

Figure 7: Timelines and take-off scenarios

##### 4.4.1.1 A Hard Take-Off

A hard take-off is shown in figure 7a. The literature refers to a short $D_2$ as a hard take-off scenario or sometimes also called a fast take-off [71]. Figure 7a shows a relatively long duration $D_1$ which translates into a slow progress on the front to create AGI.[48] This is then followed by an intelligence explosion where capabilities are developed in a very short amount of time $D_2$. Some refer to an intelligence explosion because they expect that an AI could develop its capabilities to extreme levels within minutes, hours or days [3], [72]. For elaborations on how the microeconomics of an intelligence explosion could work and why it might be likely or unlikely, see Eliezer Yudkowsky's paper [7] or Robin Hanson's and Eliezer Yudkowsky's de-

[47] Figure 7a is a simplified version of Bostrom's figure in [3, p. 63].

[48] The characterization of slow and hard take-offs is not dependent on the length of $D_1$. Only the length of $D_2$ is relevant.

bate in book format [73]. Let us assume for a moment that a hard take-off is taking place. Humans would have no time to react or to make any adjustments. The preparations that we had put in place before the AGI arrival are our only safety measures.

#### 4.4.1.2 A Soft Take-Off

Figure 7b depicts a soft take-off scenario.[49] In this specific example, the duration $D_1$ is short[50] while the take-off time to ASI is relatively long. A soft take-off is expected to happen over the duration of decades or centuries [7, p. 63]. This scenario offers humans time to steer the process of the take-off. Various approaches could be tested, new institutions could be put in place and treaties could be negotiated [3, p. 64]. In general, humanity has time to react to the increasing capabilities and their potential consequences. Precautionary measures that date back into $D_1$ would have less impact as approaches that are developed during $D_2$ would be addressing the exact problems more precisely.

#### 4.4.1.3 A Moderate Take-Off

We do not provid a graphical representation of a moderate take-off scenario. However, one can imagine some scenario that would lie between the extremes of soft and hard take-off. The literature refers here to a timeline of months or years [3, p. 64]. Humans would have limited possibilities of steering the take-off process and some but not much time to react to changes. Depending on how close a moderate take-off would be to the soft or to the hard one, there would be more or less room for influence over the take-off process, respectively.

### 4.4.2 Scenario: A Late Open Configuration

Let us now consider a concrete example scenario. Time has passed by and we are in the future at time period $t = 2075$. The openness configuration has changed from $x_{2020}$ to $x_{2075}$ as depicted in table 4.

| Dimension | $\mathbf{x_{2020}}$ | $\mathbf{x_{2075}}$ |
|---|---|---|
| Science | high | very high |
| Source Code | very low | very high |
| Data | low | very high |
| Safety Techniques | high | very high |
| Development Tools | high | very high |
| Capabilities | high | very high |
| Organizational Goals | medium | very high |
| Forecasts & Timelines | very low | very high |
| Composition of Ethic Boards | very low | very high |
| Timing of Release | high | very high |
| Distribution of Release | high | very high |

Table 4: Example configurations of the openness space for the years 2020 and 2075

Admittedly, $x_{2075}$ is an outcome as extreme as it can be. We think that slight variations would also suffice to make our point. However, we will make use of this hyperbolic scenario to make it more accessible to the reader. For our scenario, we wanted a date that would come close to the emergence of AGI. Thus, we picked the year 2075 because Müller's and Bostrom's survey showed that AI experts estimate a 90% probability for AGI to arrive before the year 2075 [45].[51] As predicted, AGI seems to be around the corner as impressive breakthroughs are justifiably[52] reported and advanced applications are flooding the market. A plethora of AI labs work on AGI while all advancements are shared openly. We assume further that this world is a software-constrained one as explained in section 3.2.2. Given that the configuration is in the extremely open "corner" of the openness space, the discounted progressing transition probabilities are strongly favoring to remain in this part of the space. The openness gradient or bias increases the probability again to stay in this corner.

### 4.4.3 Reactions and Model Adaptations

In our example scenario, a general sentiment is spreading – one that implies AGI is actually around the corner. People become more and more worrisome. Due to high to maximal openness, the AI race has been rendered increasingly competitive which exacerbated the control and political problem (see section 3.2.2). Furthermore, the acceleration of AI development complicated the control problem further. Concerns are growing and more closedness is demanded – especially in the technical openness aspects of science, source code and data. A slow down is required to safely stabilize the final sprint towards AGI. We could say that our model needs an adaptation when this happens. We could overlay the transition probabilities over all configurations with another layer of transition probabilities that point towards *very low* values for *science*, *source code* and *data*.

### 4.4.4 Possible Consequences

Even if we assume that there is a strong incentive to move away from such a high amount of openness, it might not be enough to enable a swift enough change in openness. Note that we are talking about openness of e.g. science and source code with respect to new publications. Already existing published research is hardly removable from the internet as discussed in section 4.3.5. Even if we try to look only forward, it might already not be such a simple endeavor to prevent AI researchers and engineers from publishing their new research results. In this scenario, a general development towards more openness has been happening over the duration of many years. Some researchers might have been growing up into a society which exhibits a mentality of extreme openness. Telling them suddenly in 2075 that this has been all wrong or merely

[49] Sometimes a *soft take-off scenario* is also called a *slow take-off scenario*.

[50] Again, $D_1$ is not relevant to classify a take-off scenario.

[51] Any arbitrary year in the future would have sufficed but expert predictions are probably better than nothing and thus we settled on an educated guess of 2075 to make the scenario more realistic.

[52] Compared to news that are flooded with supposed breakthroughs based on over-hype and sensational news coverage as it is often done in the late 2010s [74]–[76].

good in the past but not anymore, might encounter resistance.

Let us assume further that a hard take-off scenario is imminent. Given all these circumstances, the probability is rather high that there will be a continued rush towards the finish line – the finalization of AGI. Moreover, given that openness is so high, the probability for multipolar outcomes is increased as well which would probably imply a negative scenario for humanity (see section 3.2.2). Additionally, there is no time for humans to intervene as soon as AGI has been reached and the take-off towards ASI has commenced. The probabilities do not look rosy and a pressing existential threat is hanging over humanity. The impending scenario appears devastating.

#### 4.4.5 Admission of Uncertainties

The scenario ends abysmally – probably. However, plenty of assumptions have to be true to arrive at this final scenario. Not all of them are necessarily most probable.

First, we assume a general development into a very open part of the openness space. As mentioned before, although there might be quite some wiggle room in terms of how extremely open the configuration would be, some less extremely open configurations would lead to similar results. Second, we assume that AGI actually arrives soon from the point of 2075. Also this has not to be necessarily the case. Earlier AGI arrival but also later one could be possible as well. Earlier arrival would aggravate the problem while later arrival might alleviate it to some degree. Third, we assume that the new incentive to leave the openness corner is not high enough to leave it quickly enough. This assumption could be false as we might overestimate the influence of the discounted progressed transition probabilities over the the influence of the new incentive. However, it remains the case that already openly published research is hardly removable. Fourth, we assume that the control problem and the political problem have increasingly exacerbated over the years due to openness. But other factors next to openness could have impacted the two problems in a positive way such that the relative influence of openness might have been smaller than believed. Fifth, we assumed that the world is a software-constrained one. The overall negative impact of openness is increased if this is the case. If the world would instead be a hardware-constrained one, it might be easier to restrict big parts of the population from using powerful hardware to train or run potential AGI algorithms. Finally, we assumed a hard take-off scenario. There is no consensus about which take-off scenario is the most likely. But there seems to be consensus regarding the hard take-off being the most negative in expectation [73]. In a more moderate or even slow take-off scenario, it is more likely that there will be sufficient time to solve the control and the political problem.

Provided that only a specific combination of assumptions has to be true, there are plenty of other scenarios that look more auspicious and are potentially more likely. However, as the fate of humanity is at stake, we should not forget the high negative expected impact and we should thus steel ourselves for the worst possible scenarios. If we can reduce the probability of catastrophic scenarios even by small amounts, the overall expected impact could be drastically improved.

# 5 The Near-Term Benefit Argument

Let us zoom out and look at the big picture. For the Reconciliation Argument to be true, we need a crucial proposition to hold. This proposition is a conjunction whose first part is the following proposition:

**Definition 5.1: Near-Term Benefit Proposition**

NT has a specific level of expected value when focusing only on an NT issue and its corresponding near-term impact on society. NT's expected value is significantly higher if NT considers LT issues into their decision.

To show that this proposition holds, we introduce first two different ways of how NT and LT might prioritize their corresponding issues. Next, we make use of the insights from the last section on path dependence to demonstrate why we consider the proposition in definition 5.1 likely to be true.

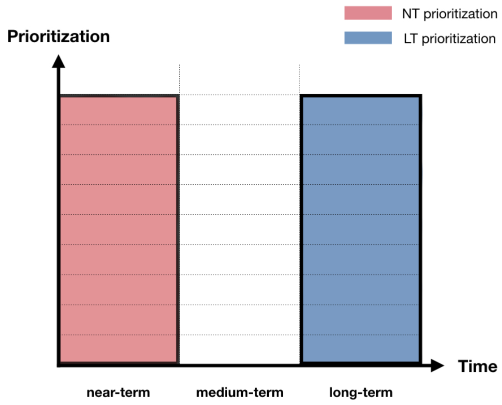
## 5.1 Different Prioritization Functions

Let us inspect the graphs in figure 8. We plot prioritization against time. Here, we partition time into three windows – near-term, medium-term and long-term. Both NT and LT have a specific amount of prioritization mass that they can distribute over different time horizons and its corresponding issues. The red mass corresponds to NT and the blue mass to LT.

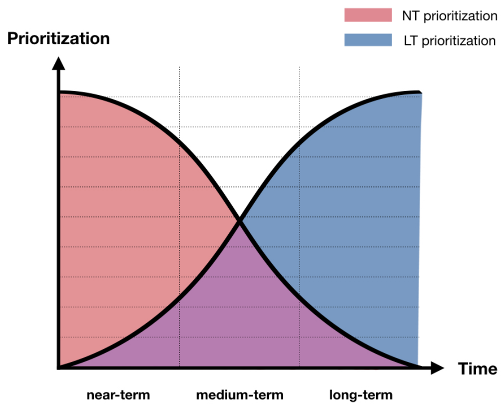
Figure 8a represents a distribution of both NT's and LT's prioritization masses in a way that might seem intuitive to some people – at least at first glance. Finding all NT prioritization mass allocated in the near-term section of the time-axis and the LT prioritization mass allocated in the long-term section is hardly a surprise. But is this the most realistic allocation of the respective prioritization masses? We do not think so as could have been guessed from the caption of figure 8b. We consider the second distribution to be a more realistic one.[53]

Let us remember the definition 2.1 of NT which states that NT ”strongly prioritize near-term issues”. The definition does not state ”exclusively” but ”strongly”. Thence, some value is also allocated to other issues that we might find somewhere in the medium-term and even in the long-term. It seems plausible to assume that we would find most NT prioritization mass in the near-term section while

[53]Figure 8b is supposed to show a qualitative difference to the distribution in figure 8a rather than a quantitative one. Consequently, a more realistic representation would contain a more asymmetric and messy distribution. For our purposes, it should suffice to deal with such a simplified distribution.




(a) Strict prioritization




(b) Realistic prioritization

Figure 8: Graphical representation of two prioritization functions

the prioritization mass becomes more and more scarce the further away we progress in time. The same holds for LT prioritization mass which agglomerates in the long-term section and continuously decreases when approximating the starting point of the near-term section. As we can tell by the purple area in figure 8b, there is some overlap between the prioritization of both NT and LT. How this relates to the prior insights of this paper, is revealed in the next subsection.

## 5.2 Avoiding Lock-ins

If we maximize the value that we could obtain in the near-term, we might lose the value that we could have attained in the medium-term and the long-term.[54]

Imagine the following world $world_{ideal}$: If the same openness configuration (e.g. $x_{near}$ in table 4) would be optimal for any term, there would be no problem. We would be able to attain the maximum value for each term and thus a maximum sum of values over all three terms. However, reality is more intricate than this. Imagine an extremely simplified world $world_{simple}$ in which it is the case that a high degree in openness in the near-term has a high NT-value, in the medium-term a medium NT-value and a low NT-value in the long-term. In this case, it would be ideal from an NT perspective if we could start out with a very high degree of openness, then decrease openness in the medium-term and further decrease it in the long-term. Given that the timing of decreasing openness would be well chosen, we would gain the highest overall NT-value as every time section would have its maximum NT-value.

However, starting with an open configuration makes it more likely to stay close to this configuration or move towards even more open configurations – that is the lock-in we were referring to in section 4.4.4. Actively changing the development direction in the openness space when being in the medium-term or in the long-term, seems less likely to be successful due to the earlier discussed discounted progressed transition probabilities and the general gradient towards more openness (see sections 4.3.4 and 4.3.5, respectively). Given these probabilities, starting open in order to collect the value of openness in the near-term, will also lead to a situation in which there is a reduced amount of value from the medium-term section (as mixed but less open is better there) and little to no value from the long-term section (as more closed would be better there). As a consequences, focusing only on the near-term benefit would imply collecting only value from the near-term section, some value from medium-term section and no value from the long-term section.

A great deal of potential value could be lost like this. If we change the openness configuration, we might lose some NT-value from the near-term section but we might gain value from the medium-term and the long-term sections. The expected NT-value can be even higher if we aim for a configuration that maximizes LT-value in the long-term rather than for a configuration that maximizes NT-value in the near-term. In appendix A, we constructed a more intricate world, argued for introducing specific numbers for our variables and calculated expected values. We compared these values and showed how maximizing for the near-term is not necessarily the best choice from an NT perspective.

Given the nature of the realistic prioritization function and our informal argument, we think that we provided sufficient evidence to believe that the Near-Term Benefit proposition is likely true.

# 6 The Long-Term Benefit Argument

Let us return again to the big picture. For the Reconciliation Argument to be true, we need a specific conjunction to hold. We have shown that its first proposition can be considered true. Now, we have to face its second proposition:

[54] In this section it will be the case that when we talk about *value* or *expected value*, we always refer to the value or expected value from the *perspective of NT*.

**Definition 6.1: Long-Term Benefit Proposition**

LT has a specific level of expected value when focusing only on an LT issue and its corresponding long-term impact on society. LT's expected value is significantly higher if LT considers NT issues into their decision.

The exact same argument will not work to show the truth of the the Long-Term Benefit Proposition. Although the realistic prioritization function in figure 8 applies to the LT prioritization as it does to the NT prioritization, the chronological order is reversed. In the former section, NT's incentive consisted always in acting on openness now in the near-term – independent from the decision whether to maximize only in the near-term or over all terms.

If there would be no risk of a lock-in and moving within the openness space would be effortless and swift, LT could start advocating for moving into more closedness at a late point in time (e.g. in the end of the medium-term or the beginning of the long-term). Action with respect to openness would not necessitate action in the near-term. Maximal value could be gained without having to meddle in the openness process before it becomes relevant for ASI in the long-term. In the meantime, LT-earmarked resources could be directed towards questions that concern exclusively the long-term such as:

- the long-term trajectories of human civilization [53]
- advanced decision theory[55]
- challenges related to transhuman and posthuman societies [79]–[82]
- foundational research on "solving ethics", i.e. figuring out what moral values an ASI should exhibit[56]

However, as we have argued in this paper, the openness process is an outcome and equilibrium dependent process which has a unique lock-in equilibrium. If LT would decide not to intervene in the openness process before it reaches the long-term time section, there might be a good chance that we end up in a configuration that is rather open. As we have shown in section 4.4, a timely escape from a configuration in this corner in the openness space is likely to be very hard to impossible. The arrival of an ASI in such a setting would be disastrous and thus the expected LT-value would be very low. In summary, if LT focus only on the long-term by researching only LT-related questions (as described above) and only attend to NT issues when they become urgent (i.e. close to long-term), then the expected LT-value is very low.

[55] E.g. *timeless decision theory* [77] and *multiverse-wide cooperation via correlated decision making* [78].

[56] Bostrom discussed several approaches such as coherent extrapolated volition [83], [84], moral rightness and moral permissibility [3, p. 209-221]

If LT weigh in on NT issues (e.g. the openness process), early and in a regular fashion, the probability is decreased that we will end up in a lock-in configuration. Not being locked-in in the long-term would enable us to either stay in a more closed subspace or to move towards it. A more closed configuration, such as $x_{long}$ in table 4 would substantially increase the probability of a successful arrival of friendly ASI. Consequently, the expected LT-value would be rather high when compared to the scenario in which LT ignores the openness process in the near-term and in the medium-term. Thus, we can consider these results as a support for the truth of the Long-Term Benefit proposition.

# 7 Conclusion

## 7.1 Uncertainties

Answering our two research questions required us to chart an openness space and to discuss how different openness configurations could affect society in the near-term and in the long-term. Furthermore, we had to model the traversing of the openness space in a dynamic way, accounting for various probability distributions and transitions. We showed why the openness process exhibits a lock-in equilibrium and what it means for the expected value for NT and for LT.

We have to concede that there is a great deal of uncertainty involved into our models and into our reasoning. The openness space might be incomplete due to missing dimensions. Some dimensions such as *science* can not easily be changed in terms of moving towards more closedness. As explained earlier, publicly funded research projects require often potential societal value. Not publishing the results of such a research project would render it useless. Allocating public funds to AI research would become hard and thus, a consequence could consist in a general diminishing of public funding for AI research. This could, however, mean that most research would be funded by corporations. This might be not desirable as not all stakeholders' interests are represented by some profit-maximizing company. Hence, decreasing the openness in one dimension might entail complications and undesired side effects. Future research could be dedicated to explore such ramifications.

Moreover, our analysis of openness on the impact on the near-term and the long-term is very likely to be not nuanced enough and probably lacking further crucial considerations. A further exploration of the main effects of openness on the control and political problem would be desirable in future research. Such research could make the picture clearer and improve the quality of our model. If it turns for example that the gradient towards openness is less straight forward than we have presented it, it could mean that moving from more to less openness on some specific dimensions is easier than on others. This could have substantial implications for how to avoid lock-ins.

Although such uncertainties remain and more research would be necessary to reduce them, we think that our argument and model is worth considering because they provide a new angle on how to think about the debate between the near-term and the long-term.

### 7.2 Reconciliation

With the help of our formal argument, we have shown that both, the Near-Term Benefit Proposition and the Long-Term Benefit Proposition are true. Thus, the answer to the first research question is a positive one – both NT and LT are reconcilable given that NT consider ASI into their decision and LT consider openness in AI development into their decision. This was done by viewing openness as a multi-dimensional space and modelling the traversing of this space as a dynamic process with shifting probability distributions over future outcomes. We identified that this openness process has the properties of being outcome and equilibrium dependent while exhibiting a unique equilibrium. Moving in the openness space from a closed configuration to an open is rather easy[57] and moving from an open configuration to a closed one is rather hard. Given enough time, a lock-in will be reached which is bad for both NT and LT. Avoiding such a lock-in, however, will benefit both factions. In some sense, we could conclude that the situation between NT and LT has been somewhat similar to the two factions in *Monty Python's Life of Brian*. The differences are negligible as the goals of those two factions are quite overlapping. A reconciled approach will render unnecessary quarrels obsolete and will thus imply that these resources can be used more effectively to face the AI risks together.

### 7.3 Action Recommendations

What action recommendations can be given to researchers and engineers in the field of AI? In expectation, it would be better to handle the following openness dimensions as openly as possible: safety techniques, organizational goals and composition of ethics boards. Being open about these aspects benefits both NT and LT directly. To avoid a lock-in, it would be expedient to reduce or at least maintain the current degree of openness in science, source code, data and development tools. The more closed these technical openness dimensions are, the more positive the effects would be on the speed of AI development and the competitiveness of the AI race and thus on the control and political problem. Given that science is exhibiting a very high degree of openness, it might be most promising to focus on not escalating the matter in the other three aspects. Especially source code is currently handled rather closed. If a general push towards more openness in the research community is not countered, especially source code will become more and more open. Later movements towards more closedness on this dimension, might be very hard as has been elaborated in this paper.

[57]This is because moving towards openness is favored by the general probability distributions over future outcomes.

### 7.4 General Lessons from Applying Path Dependence Reasoning to NT Issues

We think that it is possible to infer general lessons from applying path dependence reasoning to NT issues: If possible, we should try to model an NT issue dynamically by using general intuitions to determine probability distributions and transitions. From this model, we can infer the model's properties. These will help to discern whether the process has any equilibria and thus potential lock-ins. If it has any, we have to analyze whether they are potentially hazardous from an LT perspective. If so, we should try to recommend actions and policies that will guide us away from such lock-ins.

A possible next NT issue could be AI-augmented mass surveillance. One could imagine that more surveillance is considered highly undesired in the near-term. But it could become desired or even necessary in a later stage in time when control might necessitate to know who is working on advanced AI such as AGI or ASI.[58] At this point, we will abstain from a full analysis of this matter. But changing from a culture, which is characterized by strong anti-surveillance sentiments, to a culture that accepts or even welcomes it, seems unlikely. We could imagine that potential lock-in situations might be foreseeable. Whether such an analysis would indeed turn out fruitful, is a question that we cannot answer at this point. It remains to say that there might be insights gained to apply path dependence reasoning to other NT issues.

[58]In this context, we recommend to read *The vulnerable world hypothesis* by Nick Bostrom [85] who discussed how increasing technological progress might render the world more and more vulnerable and what we might can do about it.

# References


[1] S. Cave and S. S. ÓhÉigeartaigh, "Bridging near-and long-term concerns about ai", *Nature Machine Intelligence*, vol. 1, no. 1, p. 5, 2019.

[2] N. Bostrom, "How long before superintelligence?", *Jour. of Future Studies*, vol. 2, 1998.

[3] ——, "Superintelligence: Paths, dangers", *Strategies, Oxford University Press*, 2014.

[4] S. Baum, A. Barrett, and R. V. Yampolskiy, "Modeling and interpreting expert disagreement about artificial superintelligence", *Informatica*, vol. 41, no. 7, pp. 419–428, 2017.

[5] R. V. Yampolskiy, *Artificial superintelligence: a futuristic approach*. Chapman and Hall/CRC, 2015.

[6] E. Yudkowsky, "Artificial intelligence as a positive and negative factor in global risk", *Global catastrophic risks*, vol. 1, no. 303, p. 184, 2008.

[7] ——, "Intelligence explosion microeconomics", *Machine Intelligence Research Institute, accessed online October*, vol. 23, p. 2015, 2013.

[8] N. Bostrom, "The superintelligent will: Motivation and instrumental rationality in advanced artificial agents", *Minds and Machines*, vol. 22, no. 2, pp. 71–85, 2012.

[9] K. Crawford, "Artificial intelligence's white guy problem", *The New York Times*, vol. 25, 2016.

[10] C. Garling, *Andrew ng: Why 'deep learning' is a mandate for humans, not just machines*, [accessed 5 September 2019], 2018. [Online]. Available: `https://www.wired.com/brandlab/2015/05/andrew-ng-deep-learning-mandate-humans-not-just-machines/`.

[11] C. Williams, *Ai guru ng: Fearing a rise of killer robots is like worrying about overpopulation on mars*, [accessed 20 September 2019], 2015. [Online]. Available: `https://www.theregister.co.uk/2015/03/19/andrew_ng_baidu_ai/`.

[12] S. Castellanos, *Ai guru andrew ng on the job market of tomorrow*, [accessed 22 September 2019], 2018. [Online]. Available: `https://www.wsj.com/articles/ai-guru-andrew-ng-on-the-job-market-of-tomorrow-1540562400`.

[13] R. Brooks, *[forai] steps toward super intelligence i, how we got here*, [accessed 8 September 2019], 2018. [Online]. Available: `https://rodneybrooks.com/forai-steps-toward-super-intelligence-i-how-we-got-here/`.

[14] P. Voss, "Essentials of general intelligence: The direct path to artificial general intelligence", in *Artificial general intelligence*, Springer, 2007, pp. 131–157.

[15] L. Muehlhauser, *What is agi?*, [accessed 6 December 2019], 2013. [Online]. Available: `https://intelligence.org/2013/08/11/what-is-agi/`.

[16] R. Wiblin, *Positively shaping the development of artificial intelligence*, [accessed 6 September 2019], 2017. [Online]. Available: `https://80000hours.org/problem-profiles/positively-shaping-artificial-intelligence/`.

[17] ——, *Why the long-term future of humanity matters more than anything else, and what we should do about it*, [accessed 7 September 2019], 2017. [Online]. Available: `https://80000hours.org/podcast/episodes/why-the-long-run-future-matters-more-than-anything-else-and-what-we-should-do-about-it/`.

[18] V. Krakovna, *What are the tradeoffs between immediate and longer term ai safety efforts?*, [accessed 16 December 2019], 2018. [Online]. Available: `http://approximatelycorrect.com/2018/01/28/is-there-a-tradeoff-between-safety-concerns-about-current-and-future-ai-systems/#more-628`.

[19] T. Jones, *Monty python's life of brian*, HandMade Films & Python (Monty) Pictures.

[20] S. D. Baum, "Reconciliation between factions focused on near-term and long-term artificial intelligence", *AI & SOCIETY*, vol. 33, no. 4, pp. 565–572, 2018.

[21] J. Whittlestone, *Thoughts on short- vs. long-term ai policy*, [accessed 12 September 2019], 2019. [Online]. Available: `https://jesswhittlestone.com/blog/2019/6/22/thoughts-on-short-vs-long-term-ai-policy`.

[22] S. E. Page *et al.*, "Path dependence", *Quarterly Journal of Political Science*, vol. 1, no. 1, pp. 87–115, 2006.

[23] N. Bostrom, "Strategic implications of openness in ai development", *Global Policy*, vol. 8, no. 2, pp. 135–148, 2017.

[24] V. Krakovna, *Analysis: Clopen ai – openness in different aspects of ai development*, [accessed 10 October 2019], 2016. [Online]. Available: `https://futureoflife.org/2016/08/03/op-ed-clopen-ai-openness-in-different-aspects-of-ai-development/`.

[25] J. Whittlestone and A. Ovadya, “The tension between openness and prudence in ai research”, *arXiv preprint arXiv:1910.01170*, 2019.

[26] E. Brown and C. Woolston, “Why science blogging still matters”, *Nature*, vol. 554, no. 7690, 2018.

[27] S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach*, Second. Malaysia; Pearson Education Limited, 2014.

[28] ArXiv, *Arxiv.org*, [accessed 15 December 2019], 2019. [Online]. Available: `https://arxiv.org`.

[29] GitHub, *Github*, [accessed 15 December 2019], 2019. [Online]. Available: `http://www.github.com`.

[30] Nature, *Reporting standards and availability of data, materials, code and protocols*, [accessed 17 October 2019], 2019. [Online]. Available: `https://www.nature.com/nature-research/editorial-policies/reporting-standards#availability-of-computer-code`.

[31] OpenAI, *Gpt-2: 1.5b release*, [accessed 11 October 2019], 2019. [Online]. Available: `https://openai.com/blog/gpt-2-1-5b-release/`.

[32] K. Johnson, *Nvidia trains world’s largest transformer-based language model*, [accessed 15 October 2019], 2019. [Online]. Available: `https://venturebeat.com/2019/08/13/nvidia-trains-worlds-largest-transformer-based-language-model/`.

[33] A. Rathi, *“ai sells, data delivers!” importance of data in the age of fancy ai algorithms*, [accessed 16 December 2019], 2019. [Online]. Available: `https://towardsdatascience.com/ai-sells-data-delivers-34dc528c8533`.

[34] Kaggle, *Datasets*, [accessed 15 December 2019], 2019. [Online]. Available: `https://www.kaggle.com/datasets`.

[35] OpenAI, *Resources*, [accessed 16 October 2019], 2019. [Online]. Available: `https://openai.com/resources/`.

[36] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, “Concrete problems in ai safety”, *arXiv preprint arXiv:1606.06565*, 2016.

[37] C. Wainwright and P. Eckersley, *Introducing safelife: Safety benchmarks for reinforcement learning*, [accessed 6 December 2019], 2019. [Online]. Available: `https://www.partnershiponai.org/safelife/`.

[38] D. Clark, *Top 16 open source deep learning libraries and platforms*, [accessed 16 December 2019], 2018. [Online]. Available: `https://www.kdnuggets.com/2018/04/top-16-open-source-deep-learning-libraries.html`.

[39] D. Rowan, *Astro teller of google[x] wants to improve the world’s broken industries*, [accessed 6 December 2019], 2013. [Online]. Available: `https://www.wired.co.uk/article/destination-moon`.

[40] Google, *About google*, [accessed 20 October 2019], 2019. [Online]. Available: `https://about.google/`.

[41] OpenAI, *About openai*, [accessed 20 October 2019], 2019. [Online]. Available: `https://openai.com/about/`.

[42] J. Vincent, *The problem with ai ethics*, [accessed 10 November 2019], 2019. [Online]. Available: `https://openai.com/about/`.

[43] S. Shead, *Deepmind’s mysterious ethics board will reportedly ’control’ agi if it’s ever created*, [accessed 18 November 2019], 2019. [Online]. Available: `https://www.forbes.com/sites/samshead/2019/03/14/deepminds-mysterious-ethics-board-will-reportedly-control-agi-if-its-ever-created/#754cc1e852a9`.

[44] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, and J. Wang, “Release strategies and the social impacts of language models”, *arXiv preprint arXiv:1908.09203*, 2019.

[45] V. C. Müller and N. Bostrom, “Future progress in artificial intelligence: A survey of expert opinion”, in *Fundamental issues of artificial intelligence*, Springer, 2016, pp. 555–572.

[46] S. Shead, *Facebook reportedly has a dedicated ai ethics team*, [accessed 18 November 2019], 2018. [Online]. Available: `https://www.forbes.com/sites/samshead/2018/05/03/facebook-reportedly-has-a-dedicated-ai-ethics-team/`.

[47] K. Piper, *Exclusive: Google cancels ai ethics board in response to outcry*, [accessed 18 November 2019], 2019. [Online]. Available: `https://www.vox.com/future-perfect/2019/4/4/18295933/google-cancels-ai-ethics-board`.

[48] A. Boyle, *Microsoft is turning down some sales over ai ethics, top researcher eric horvitz says*, [accessed 16 December 2019], 2018. [Online]. Available: `https://www.geekwire.com/2018/microsoft-cutting-off-sales-ai-ethics-top-researcher-eric-horvitz-says/`.

[49] G. Sachs, *The real consequences of artificial intelligence*, [accessed 18 November 2019], 2015. [Online]. Available: `http://www.cognitivescale.com/wp-content/uploads/2015/03/FT-Artificial-Intelligence.pdf`.

[50] Forbes, *Ai and healthcare: A giant opportunity*, [accessed 18 November 2019], 2019. [Online]. Available: https://www.forbes.com/sites/insights-intelai/2019/02/11/ai-and-healthcare-a-giant-opportunity/#3c345284c682.

[51] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries", *arXiv preprint arXiv:1905.08883*, 2019.

[52] A. Radford, J. Wu, D. Amodei, D. Amodei, J. Clark, M. Brundage, and I. Sutskever, "Better language models and their implications", *OpenAI blog*, 2019.

[53] S. D. Baum, S. Armstrong, T. Ekenstedt, O. Häggström, R. Hanson, K. Kuhlemann, M. M. Maas, J. D. Miller, M. Salmela, A. Sandberg, *et al.*, "Long-term trajectories of human civilization", *foresight*, vol. 21, no. 1, pp. 53–83, 2019.

[54] R. Hanson, "If uploads come first", 1994.

[55] N. Bostrom, "The future of human evolution", *Death and anti-death: Two hundred years after Kant, fifty years after Turing*, pp. 339–371, 2004.

[56] R. Hanson, "Employment in the age of em: Simulated brains and the economics of labor", in *Surviving the Machine Age*, Springer, 2017, pp. 51–62.

[57] L. Mallinson, J. Russell, D. D. Cameron, J. Ton, P. Horton, and M. E. Barker, "Why rational argument fails the genetic modification (gm) debate", *Food Security*, vol. 10, no. 5, pp. 1145–1161, 2018.

[58] S. Liebowitz, S. Margolis, *et al.*, "Encyclopedia of law and economics", *E Cheltenham*, 2000.

[59] C. Crouch and H. Farrell, "Breaking the path of institutional development? alternatives to the new determinism", *Rationality and society*, vol. 16, no. 1, pp. 5–43, 2004.

[60] P. Pierson, *Politics in time: History, institutions, and social analysis*. Princeton University Press, 2011.

[61] S. J. Liebowitz and S. E. Margolis, "Path dependence", *Encyclopedia of law and economics*, 1999.

[62] J. Torfing, "Rethinking path dependence in public policy research", *Critical Policy Studies*, vol. 3, no. 1, pp. 70–83, 2009.

[63] P. A. David, "Path dependence, its critics and the quest for 'historical economics'", *Evolution and path dependence in economic ideas: Past and present*, vol. 15, p. 40, 2001.

[64] B. Valley, "Entrenched in power: Path dependence and un security council reform", 2015.

[65] Wikipedia, *Qwerty*, [accessed 26 November 2019], 2006. [Online]. Available: https://en.wikipedia.org/wiki/QWERTY#/media/File:KB_United_States.svg.

[66] ——, *Dvorak keyboard layout*, [accessed 26 November 2019], 2009. [Online]. Available: https://en.wikipedia.org/wiki/Dvorak_keyboard_layout#/media/File:KB_United_States_Dvorak.svg.

[67] W. B. Arthur, *Increasing returns and path dependence in the economy*. University of michigan Press, 1994.

[68] G. B. de la Fuente, *What is open science? introduction*, [accessed 04 December 2019], 2019. [Online]. Available: https://www.fosteropenscience.eu/content/what-open-science-introduction.

[69] M. Möbus, *Openai finally releases "dangerous" large-scale unsupervised language model gpt-2*, [accessed 04 December 2019], 2019. [Online]. Available: https://jaxenter.com/openai-releases-language-model-gpt-2-163881.html.

[70] M. E. Roberts, *Censored: distraction and diversion inside China's Great Firewall*. Princeton University Press, 2018.

[71] E. Yudkowsky, *Hard takeoff*, [accessed 7 December 2019], 2008. [Online]. Available: https://www.lesswrong.com/posts/tjH8XPxAnr6JRbh7k/hard-takeoff.

[72] L. Muehlhauser, *Intelligence explosion faq*, [accessed 7 December 2019], 2013. [Online]. Available: https://intelligence.org/files/IE-FAQ.pdf.

[73] R. Hanson, "The hanson-yudkowsky ai-foom debate", *Berkeley, CA: Machine Intelligence Research Institute*, 2013.

[74] G. Marcus and E. Davis, *Rebooting AI: Building Artificial Intelligence We Can Trust*. Pantheon, 2019.

[75] P. Cohan, *3 reasons ai is way overhyped*, [accessed 7 December 2019], 2019. [Online]. Available: https://www.forbes.com/sites/petercohan/2019/02/15/3-reasons-ai-is-way-overhyped/#39a5e4d75a6a.

[76] O. Schwartz, *'the discourse is unhinged': How the media gets ai alarmingly wrong*, [accessed 7 December 2019], 2018. [Online]. Available: https://www.theguardian.com/technology/2018/jul/25/ai-artificial-intelligence-social-media-bots-wrong.

[77] E. Yudkowsky, "Timeless decision theory", *The Singularity Institute, San Francisco*, 2010.

[78] C. Oesterheld, "Multiverse-wide cooperation via correlated decision making", *Foundational Research Institute*, 2017, [accessed 17 December 2019]. [Online]. Available: https://foundational-research.%20org/multiverse-wide-cooperation-via-correlated-decision-making.

[79] N. Bostrom, "In defense of posthuman dignity", *Bioethics*, vol. 19, no. 3, pp. 202–214, 2005.

[80] D. Roden, *Posthuman life: Philosophy at the edge of the human*. Routledge, 2014.

[81] N. Bostrom *et al.*, "Dignity and enhancement", *Contemporary Readings in Law and Social Justice*, vol. 1, no. 2, pp. 84–115, 2009.

[82] P. K. Nayar, *Posthumanism*. John Wiley & Sons, 2018.

[83] E. Yudkowsky, "Coherent extrapolated volition", *Singularity Institute for Artificial Intelligence*, 2004.

[84] N. Tarleton, "Coherent extrapolated volition: A meta-level approach to machine ethics", *Machine Intelligence Research Institute*, 2010.

[85] N. Bostrom, "The vulnerable world hypothesis", *Global Policy*, 2018.

# Appendix

## A Support For the Near-Term Benefit Argument: Example Calculations of Expected Values for NT

In this section, we first set up the world that we assume we live in. In the subsequent sections we calculate expected values with two different weight vectors over the time sections.

### A.1 World Setting

In the subsection 5.2, we introduced $world_{ideal}$ and $world_{simple}$ which are both far from our real world. Now, we want to introduce a more realistic world which we call $world_{real}$. This world implies that we do not look at the overall openness but at the subtle changes in openness aspects which can be interpreted as higher or lower in value from an NT or an LT perspective. Table 5 provides potentially ideal configurations of the openness space from these two perspectives. Configuration $x_{near}$ corresponds to the configuration that maximizes the NT-value in the near-term section. Configuration $x_{long}$ refers to the configuration that maximizes the NT-value in the long-term section. We call them *ideal* configurations while we advise to take our word with a grain of salt. The configurations were chosen on the basis of our analysis in section 3.2 which is an approximation of the corresponding impact but suspect to incompleteness and errors. Additionally, due to ambiguous impacts of some aspects, we settled on assigning such aspects *medium* values. It is unclear to us whether this is sufficient or whether it would be better to distinguish various configurations where we explore value variations in the ambiguous aspects. As this detailed exploration might however go beyond the scope of this paper, we settle for now on this simple way of dealing with ambiguous impacts of openness aspects.

Let us have a closer look at table 5 and its two ideal configurations. We have seen such a table already twice. The values of the two configurations are chosen based on our earlier analysis. While $x_{near}$ has almost only values of *very high*, $x_{long}$ is more of a mixed bag. The table has a color code. If a row is highlighted in green, it means that the ideal values for this dimension are congruent. Is a row highlighted in red, then the values are as far apart as they can be. Not highlighted rows represent dimensions whose values are somewhere in the middle – in this concrete example, they represent a distance between *very low* and *medium* or between *medium* and *very high*.

| **Dimension** | $\mathbf{x_{near}}$ | $\mathbf{x_{long}}$ |
|---|---|---|
| Science | very high | very low |
| Source Code | very high | very low |
| Data | very high | very low |
| Safety Techniques | very high | very high |
| Development Tools | very high | very low |
| Capabilities | very high | medium |
| Organizational Goals | very high | very high |
| Forecasts & Timelines | very high | medium |
| Composition of Ethic Boards | very high | very high |
| Timing of Release | medium | very low |
| Distribution of Release | very high | medium |

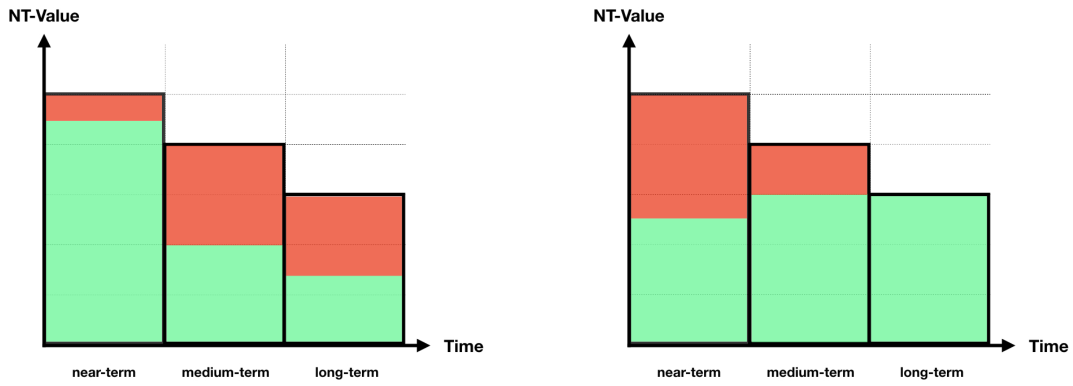
Table 5: Potentially ideal configurations of the openness space from two perspectives

### A.2 Expected Value with $W_P = (5, 4, 3)$

Now, we try to translate these configurations into expected values from an NT perspective. First, we show a graphical representation of these expected values and explain how the numbers come into existence. Then, we present some calculations.

Figure 9 provides two subfigures. We plot NT-value against time. Time is again partitioned into three sections: near-term, medium-term and long-term. The height of the bars correspond to how much NT prioritize a specific time section.[59] In figure 9, the priorities have the given weights $w_P = (5, 4, 3)$. The sum of the green colored areas of the bars refer to the NT-value. The sum of the red areas correspond to the lost value.

[59]This should be reminiscent of figure 8b whose function, however, is a continuous one and calculating the area under the curve would require cumbersome integral calculus. To simplify our calculations, we use discrete rectangles.




(a) NT expected value of $x_{near}$ (b) NT expected value of $x_{long}$

Figure 9: NT expected values given $W_P = (5, 4, 3)$

Let us turn to the question of how we determined the ratio of the colored areas in figure 9. First, note that we assume that all openness dimensions have the same weight in this model which implies that they are equally important. In reality, this is not necessarily the case but we go with this assumption for the remainder of this paper. We will now analyze the graph in figure 9a with the given configuration $x_{near}$. The first bar is mostly green due to the overall analysis of the expected near-term impact in section 3.2.1. An almost entirely open configuration would benefit the near-term substantially. We subtracted 10% on the basis of some exceptional technologies being potentially used to intentionally or accidentally harm society.[60] The green area of the second bar in figure 9a is at around 50% due to the ambiguous results in Bostrom's analysis [23]. The bar in the long-term section has a green area that is around 45% of the entire bar. To understand this choice, we have to look back at table 5 and count how many green, red and color-less rows there are. Each green dimension is worth 1 point, each red dimension 0 points and each color-less dimension is worth 0.5 points. Consequently, $3 \cdot 1 + 4 \cdot 0 + 4 \cdot 0.5 = 5$. As the maximum amount of points is 11, we get a ratio of green area to entire box at $\frac{5}{11} \approx 0.45$.

In summary, we have a relative value distribution with $V_{near} = (0.9, 0.5, 0.45)$ and prioritization weights $W_P = (5, 4, 3)$. To calculate the expected NT-value $EV_{near}$, we need a dot product, such that $EV_{near} = V_{near} \cdot W_P$. Thus,

$$EV_{near} = 0.9 \cdot 5 + 0.5 \cdot 4 + 0.45 \cdot 3 = 7.85$$

The numbers for figure 9b have been determined in a similar fashion. The bar in the near-term section has a green part of around 50%. 45% come from the same calculation as for the long-term bar in figure 9a. We added 5% as the potential risks from military uses, applications for social control and systemic risks from more reliance on complex autonomous processes are less evident in the configuration $x_{long}$. The medium-term bar is still a mixed bag, however the increased closedness in the technical openness aspects are a measure of caution which leads to a green part of 75%. The long-term bar is completely green.[61]

Thus, we have a relative value distribution $V_{long} = (0.5, 0.75, 1.0)$ and prioritization weights $W_P = (5, 4, 3)$. Their dot product results in

$$EV_{long} = 0.5 \cdot 5 + 0.75 \cdot 4 + 1.0 \cdot 3 = 8.5$$

This implies an increase in $EV$ of $\frac{EV_{long} - EV_{near}}{EV_{near}} \approx 0.083$. Hence, if we do not focus on maximizing only the expected value over the near-term section (i.e. preferring $x_{near}$) but try to maximize over all time sections, the configuration $x_{long}$ becomes even more attractive than $x_{near}$ from an NT perspective.

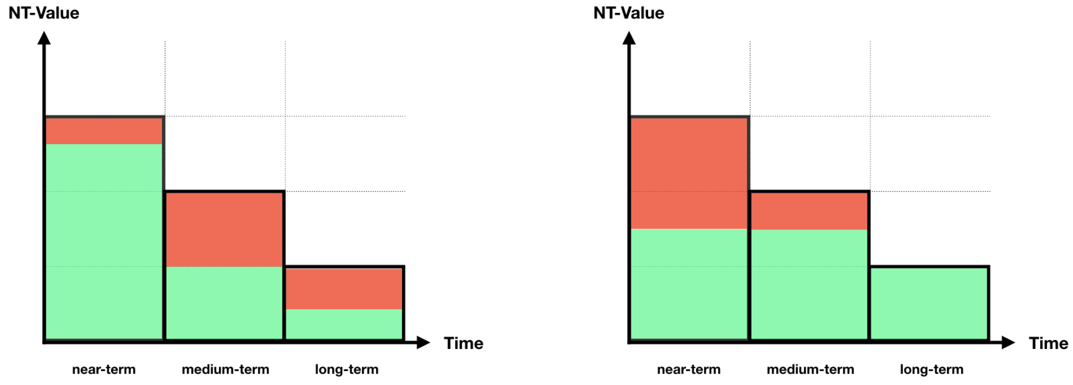
## A.3 Expected Value with $W_P = (3, 2, 1)$

The NT-value is not *necessarily* bigger for $x_{long}$ than for $x_{near}$. In the example of the previous section, it is indeed the case. If we change, however, some of the variables, we attain other results. In order to demonstrate this, we change only the prioritization weights to $W_P = (3, 2, 1)$ which means that the relative weight of the near-term value is now higher than in the former scenario. The relative weights of the former scenario were $W_{rP} = (0.42, 0.33, 0.25)$ while the relative weights of the new scenario are $W_{rP} = (0.5, 0.33, 0.17)$. Now, the weights are more skewed towards the near-term section. Figure

[60] E.g. military uses, applications for social control and systemic risks from more reliance on complex autonomous processes

[61] Again, keep in mind that this is an approximation of ideal configurations. A better, more fine-grained model would result in variations of the ideal configurations.

10 represents this new distribution of the weights. The height of the boxes is adjusted accordingly.




(a) NT expected value of $x_{near}$ (b) NT expected value of $x_{long}$

Figure 10: NT expected values given $W_P = (3, 2, 1)$

With the new prioritization weights $W_P = (3, 2, 1)$ and the old $V_{near} = (0.9, 0.5, 0.45)$ and $V_{long} = (0.5, 0.75, 1.0)$, we can calculate $EV_{near}$ and $EV_{long}$.

$$EV_{near} = 0.9 \cdot 3 + 0.5 \cdot 2 + 0.45 \cdot 1 = 4.15$$

$$EV_{long} = 0.5 \cdot 3 + 0.75 \cdot 2 + 1.0 \cdot 1 = 4$$

$$\frac{EV_{long} - EV_{near}}{EV_{near}} \approx -0.036$$

The expected NT-value of the configuration $x_{near}$ is thus higher than the expected NT-value of $x_{long}$. Other weights vectors might lead to other results. In the next section, we elaborate on the problematic nature of these findings and the significance of our results.

## A.4 Stumbling Blocks and Significance

There a plenty of uncertainties involved in our considerations and calculations.

### A.4.1 A Problem with Ideal Configurations

First, as already mentioned, we are not very certain about the *ideal* configurations in table 4. A better model or analysis would be likely to change these configurations and by transition the corresponding expected values. Assigning *medium* values to openness aspects whose impact is ambiguous, might be a bad model setting. Furthermore, giving all openness aspects the same weight is simple but questionable. More research would be necessary to determine better weights.

### A.4.2 A Problem with the Relative Value Distribution

Second, even if these before-mentioned aspects would all be certain, $V_{near}$ and $V_{long}$ could have been determined wrongly. There is some wiggle room, in how one could choose the ratio of green to red areas within the corresponding bars. Variations in these variables might be possible and thus also change the final expected values.

### A.4.3 A Problem with the Prioritization Weights

Third, the distribution of $W_P$ might be skewed in a different manner such that it is favoring the near-term section more. Given our two calculations, we could potentially assume that if the distribution $W_P$ is more equally distributed, the higher is the NT-value for configuration $x_{long}$. The obvious question is now: What kind of distribution is the correct one? The results of our calculations depend significantly on the answer to this question. As we are talking about the prioritization weights from the NT perspective, it might be useful to conduct a survey and ask people that would prioritize near-term issues over other issues, how much exactly they would prioritize near-term issues.[62] These results could then be used to be fed into our model. For the moment, we find the weights $W_P = (5, 4, 3)$ more likely to be the case than e.g. $W_P = (3, 2, 1)$.

[62] A participant of the survey could be given a prioritization mass of 100. The participant is then supposed to distribute these 100 points over the near-term, the medium-term and the long-term.

The debate is usually between NT and LT, focusing solely on the near-term and the long-term. The medium-term is not much discussed. We think that the near-term and the medium-term are usually thrown together. If this is the case, an adjusted distribution could look like this: $W_P = (9, 3)$[63] which corresponds to relative weights $W_{rP} = (0.75, 0.25)$. This seems like a realistic estimation. The model would, however, benefit from more empirical data.

### A.4.4 Configurations Between Ideal Near-Term and Ideal Long-Term Configurations

In our analysis, we have only considered two possible configurations – $x_{near}$ and $x_{long}$. However, it might be possible that a configuration, somewhere between these two, could increase the expected NT-value even more. There is potential here to gain more insights. It might just be a bit too much to include this into this paper as well.

### A.4.5 Significance for the Argument

Given our provisory results, we are tempted to say that we have shown that it is possible that even a configuration $x_{long}$ which is optimized for the long-term section, can be higher in NT-value than a configuration $x_{near}$. We suspect that there might be configurations between $x_{near}$ and $x_{long}$ that yield an even higher expected NT-value over all terms. But even with our result of $\Delta EV = 0.083$ in section A.2, we could call this somewhat a significant boost in $EV$. And thus, our results support the proposition from definition 5.1 to be true.

[63]Summing the values from near-term (5) and medium-term (4) to get to the value 9.