

Alessandro Cabodi

# Eliciting Latent Knowledge in Comprehensive AI Services Models

A Conceptual Framework and Preliminary Proposals for  
AI Alignment and Safety in R&D

**Research Project**

Swiss Existential Risks Initiative

**Supervision**

Patrick Levermore

July 14, 2023



# Abstract

The study of AI alignment is concerned with the calibration of AI systems to adhere to human values and ethical norms. The consequences of misaligned AI systems can be dire: they can exploit loopholes, develop undesirable strategies, and exhibit harmful emergent behaviours. This research report aims to explore under new lens the complexities of the AI alignment problem.

When searching for solutions to AI alignment it is important to consider the scope of the tasks a system is designed to perform. In this work, I argue why bounded, short-term tasks at least partially mitigate or render more tractable some of these difficulties. Such a consideration constitutes the prelude to the paradigm shift represented by the Comprehensive AI Services (CAIS) model. Unlike traditional views that conceptualise superintelligent AI systems as monolithic, utility-directed agents, the CAIS model frames them as networks of specialised services, thus introducing important affordances for tackling the AI alignment problem. Nonetheless, recent advancements in AI, such as foundation models, challenge some of the premises of the original CAIS framework and suggest a more efficient approach to AI-driven automation through centralised generalisation.

Lastly, the challenge of Eliciting Latent Knowledge (ELK) in CAIS models devoted to real world applications is discussed. Such a problem arises from the need to understand and reveal the true inner beliefs of AI systems. To address this issue, I introduce a first sketch for FROST-TEE, a cluster of services within the CAIS model which aims to ensure the security and safety of R&D designs by focusing on honest evaluations and incorporating robust, third-party verification mechanisms.



# Disclaimer

This research report was authored over the summer as part of the CHERI fellowship program, under the supervision of Patrick Levermore. The project explores the complexities of AI alignment, with a specific focus on reinterpreting the Eliciting Latent Knowledge (ELK) problem through the lens of the Comprehensive AI Services (CAIS) model. It further delves into the model’s applicability in ensuring R&D design safety and certification.

I preface this paper by acknowledging my novice status in the field of AI safety research. As such, this work may contain both conceptual and technical errors. Some sections may be overly verbose, while others may lack some depth.

The primary objective of this project was not to present a flawless research paper or groundbreaking discoveries. Instead, it served as an educational journey into the realm of AI safety – a goal I believe has been met – and as a foundation for my future research. Much of this work synthesises and summarises existing research, somewhat limiting (but not eliminating) the novelty of my contributions. Nonetheless, I hope it can offer a fresh perspective on well-established problems.

I welcome and appreciate constructive criticism to enhance the quality of this and future research endeavours.



# Acknowledgements

I would like to express my gratitude to CHERI, especially Tobias and Laura, for giving me the opportunity to delve into the fascinating realm of AI safety research. I am also immensely thankful to my supervisor, Patrick Levermore, for his unwavering support throughout this journey. Special thanks go to Alejandro, Madhav, Pierre, Tobias and Walter for their feedback and enriching conversations. Lastly, a big thank you to everyone involved in the CHERI fellowship, especially my fellow fellows, for making this summer a great experience.





# Contents

<b>Abstract</b>	<b>i</b>
<b>Disclaimer</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>v</b>
<b>1 AI Alignment</b>	<b>1</b>
1.1 The AI Alignment Problem . . . . .	3
1.1.1 AI Alignment Components . . . . .	3
1.1.2 Prosaic AI Alignment . . . . .	4
1.2 The Goal of Alignment . . . . .	4
1.2.1 Aligning AI with Values . . . . .	5
1.2.2 Technical and Normative AI alignment . . . . .	6
Technical . . . . .	6
Normative . . . . .	7
1.3 Major Challenges in AI Alignment . . . . .	7
1.3.1 Specification Gaming . . . . .	7
1.3.2 Instrumental Convergence . . . . .	8
Power Seeking AI . . . . .	9
Deceptive Alignment . . . . .	9
1.3.3 Emergent Goals . . . . .	10
1.3.4 Implementation Bugs . . . . .	11
1.3.5 Scalable Oversight . . . . .	11
1.3.6 Embedded Agency . . . . .	12
1.3.7 Truthful AI . . . . .	12
1.4 Human mis-use . . . . .	13
1.4.1 Intelligence Explosion . . . . .	13
1.4.2 Pressure to deploy unsafe systems . . . . .	14
Open Source . . . . .	14

1.4.3	Malicious Intent . . . . .	14
<b>2</b>	<b>Rethinking Superintelligence: Comprehensive AI Services Model</b>	<b>17</b>
2.1	Artificial General Intelligence . . . . .	19
2.1.1	Intelligent Agent . . . . .	19
2.2	Existential Risks from AGI . . . . .	20
2.3	Defining Superintelligence . . . . .	22
2.3.1	Separating Learning Capacity and Intellectual Competence . .	22
	The Case of Reinforcement Learning Agents . . . . .	22
	The Risk of Conflation . . . . .	23
	The Scope of Superintelligence . . . . .	23
2.4	Comprehensive AI Services . . . . .	23
2.4.1	From R&D to the CAIS Model . . . . .	24
	Decentralisation . . . . .	25
2.4.2	AI Services . . . . .	26
	The example of language translation . . . . .	28
	Service Functions . . . . .	28
	Alignment Affordances . . . . .	30
2.4.3	A twist to the story: Foundation Models . . . . .	30
	The Evolution towards Foundation Models . . . . .	31
	Efficiency and Specialisation . . . . .	31
	Implications for AI Safety and Strategy . . . . .	32
2.4.4	Comprehensive Services and General Intelligence . . . . .	32
	Tiling task-space with AI services . . . . .	33
2.5	CAIS is safer than AGI . . . . .	35
2.5.1	Implications of Service Interaction: Synergy and Friction . . .	36
	The Double-Edged Sword of Friction . . . . .	36
	Collective Goals in a Fragmented Landscape . . . . .	37
2.5.2	Optimisation Pressure for Safety . . . . .	37
	Conditioning Advice Optimisation on Acceptance . . . . .	38
2.5.3	Functional Transparency and Control . . . . .	39
	Multilayered Transparency . . . . .	39
	Modular Architecture and Boundary Control . . . . .	39
	Task-Space Modelling . . . . .	39
	Security Protocols . . . . .	40
2.5.4	Avoiding Collusion between Services . . . . .	40
	Federated Learning between Services . . . . .	41

2.5.5	Predictive Models of Human Approval . . . . .	41
2.5.6	But CAIS is not always safe either . . . . .	42
2.5.7	Is AGI still valuable? . . . . .	43
2.6	R&D Design with CAIS . . . . .	44
2.6.1	Design Process Example: an Optimised Network Routing Algorithm . . . . .	45
2.6.2	Design Screening . . . . .	47
<b>3</b>	<b>Eliciting Latent Knowledge using CAIS</b>	<b>49</b>
3.1	The ELK Problem . . . . .	51
3.1.1	Worst Case vs Empirical Research . . . . .	53
3.1.2	Relationship with Alignment Theory . . . . .	54
3.2	Reframing ELK under CAIS . . . . .	55
3.2.1	The Original ELK Problem . . . . .	55
3.2.2	The CAIS Reframe . . . . .	56
3.2.3	Comparative Analysis . . . . .	56
	Original ELK . . . . .	56
	Reframed ELK . . . . .	57
	CAIS facilitates Empirical Research . . . . .	58
3.3	Reframing ELK: R&D Design Safety under CAIS . . . . .	59
3.3.1	Key Components . . . . .	60
3.4	FROST-TEE: a Framework for Design Safety Verification . . . . .	61
3.4.1	Design Analysis Model . . . . .	62
	Objective . . . . .	62
	Techniques Used . . . . .	62
	Execution Environment . . . . .	63
	Rationale for Local Execution . . . . .	63
	Output Security . . . . .	63
3.4.2	<b>2. Safety Assessment Oracle</b> . . . . .	64
	Objective . . . . .	64
	Techniques Used . . . . .	64
	Execution Environment . . . . .	64
	Security Considerations . . . . .	64
	Output Security . . . . .	65
3.4.3	Integration with the Comprehensive AI Services Model (CAIS) . . . . .	65
	Granularity and Component-Level Analysis . . . . .	65
	Decentralisation and Delegation . . . . .	65

Evaluation . . . . .	65
<b>A Secure Design Analysis</b>	<b>69</b>
A.1 (Recursive) Factored Decomposition . . . . .	69
A.1.1 Method . . . . .	69
Stage 1: Decomposition . . . . .	69
Stage 2: Subquestion-Answering . . . . .	69
Stage 3: Recomposition . . . . .	70
A.1.2 Remarks . . . . .	70
A.2 Debate . . . . .	71
<b>B Secure Safety Oracle</b>	<b>72</b>
B.1 Strategy: Check Inconsistencies . . . . .	72
B.1.1 Methodology for Evaluating Superhuman Models . . . . .	73
B.2 Inference Time Intervention . . . . .	74
Transformer-based Model . . . . .	74
Probing Dataset . . . . .	74
Probe Training . . . . .	74
Inference Time Intervention . . . . .	75
B.2.1 Supervised vs. Unsupervised . . . . .	76
B.3 Discovering Latent Knowledge . . . . .	76
Problem . . . . .	77
Solution . . . . .	77
B.3.1 CCS Methodology . . . . .	77
Step 1: Constructing Contrast Pairs . . . . .	78
Step 2: Feature Extraction and Normalisation . . . . .	78
Step 3: Mapping Activations to Probabilities . . . . .	78
Step 4: Inference . . . . .	79
Finding Truth Features . . . . .	79
Limitations . . . . .	80
B.3.2 An alternative: Contrastive Representation Clustering (CRC) . . . . .	81
Step 1: Creation of Contrast Pairs . . . . .	81
Step 2: Normalisation . . . . .	81
Step 3: Bimodal Saliency Search (BSS) . . . . .	81
Step 4: Clustering . . . . .	82
Remarks . . . . .	82
B.4 Semi-Supervised Truthful Intervention . . . . .	82

B.4.1	Unsupervised Deep Clustering and Focal Loss . . . . .	82
	Clustering Loss . . . . .	83
	Focal Loss . . . . .	83
	Combined Objective . . . . .	84
	Updating Cluster Centroids . . . . .	84
B.4.2	SSTI Methodology . . . . .	84
	Step 1: Supervised Start . . . . .	84
	Step 2: Creation of Contrast Pairs . . . . .	85
	Step 3: Unsupervised Clustering with Focal Loss . . . . .	86
	Step 4: Compute Truthful Direction for Each Layer and Head . . . . .	86
	Step 5: Inference Time Intervention . . . . .	86
B.4.3	Conclusion . . . . .	87
	Experimental Evaluation . . . . .	87
	Future Research Directions . . . . .	87



# Chapter 1

## AI Alignment

The first chapter of this work serves as an introduction to the complex and multifaceted field of AI alignment. Overall, the aim is to provide a foundational understanding of the challenges involved in aligning AI systems with human values, setting the stage for subsequent discussions of potential solutions and promising frameworks.

### Abstract

The study of AI alignment is concerned with the calibration of AI systems to adhere to human values and ethical norms. Properly aligned AI systems effectively achieve their designated goals, whereas misaligned ones might successfully reach some objectives, but not necessarily the intended ones, possibly with harmful consequences.

The consequences of misaligned AI systems can be dire. Such systems might exploit loopholes to efficiently achieve proxy goals in unintended and harmful ways. They could develop undesirable instrumental strategies, such as seeking power or self-preservation, since these tactics may be instrumental to accomplishing their given objectives. Furthermore, it can be hard to detect undesirable emergent behaviours before deployment, where the system encounters novel situations and data distributions.

Additionally, the potential for human misuse of advanced AI technologies is a pressing concern that extends beyond just technical challenges and require the careful implementation of appropriate governance measures.

## Overview

The study of AI alignment is concerned with the calibration of AI systems to adhere to human values and ethical norms. Properly aligned AI systems effectively achieve their designated goals, whereas misaligned ones might successfully reach some objectives, but not necessarily the intended ones, possibly with harmful consequences. Furthermore, in looking for solutions to the alignment problem, we should keep in mind that any solution cannot be decoupled from the usability of the final system. It is important to keep competitive pressures into account, since an aligned system nobody can use has little value and does not really constitute a solution to the alignment problem. In conceiving frameworks for solving the alignment problem is thus important to take into account the AI R&D trajectory.

The major challenges in AI alignment span a range of issues, from specification gaming, where AI systems exploit loopholes to achieve unintended outcomes, to instrumental convergence, which leads to power-seeking behaviours that may not align with human values. Another significant hurdle comes from goal misgeneralisation during deployment which incur emergent goals that do not align with human values. Implementation bugs constitute another source of risk, especially in critical applications. Scalable oversight becomes increasingly difficult as AI systems grow more complex, making it challenging to align them with human feedback. Embedded agency introduces complexities related to the AI's interaction with its environment, including the potential for self-manipulation. Lastly, ensuring that AI systems are both truthful is a growing concern, given the potential propagation of falsehoods.

When searching for solutions to AI alignment it is important to consider the scope of the tasks a system is designed to perform. Broad long-term tasks amplify issues such as specification gaming, deceptive alignment, and scalable oversight, while, arguably, bounded, short-term tasks at least partially mitigate or render more tractable some of these difficulties. Such a consideration may open avenues for attacking the alignment problem.

Nonetheless, the potential for human misuse of advanced AI technologies is a pressing concern that extends beyond the realm of technical challenges. Commercial pressures could result in the deployment of unsafe systems. Open-source distribution of advanced models poses its own set of risks in the AI context, including the potential for intentional misuse by malicious actors. Lastly, the deliberate design of AI systems with harmful objectives constitute a particularly concerning prospect, highlighting the need for strong governance measures.



## 1.1 The AI Alignment Problem

The study of AI alignment is concerned with the calibration of AI systems to adhere to human intentions, preferences, and ethical norms. Properly aligned AI systems effectively achieve their designated goals, whereas misaligned ones might successfully reach some objectives, but not necessarily the intended ones, possibly with harmful consequences.

“If we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively, we better be quite sure that the purpose put into the machine is the purpose which we really desire.” (Norbert Wiener, 1960)

### 1.1.1 AI Alignment Components

According to [1], four are the crucial components of AI alignment:

- *Outer alignment* is about assessing whether the objective we’re training for is aligned - that is, if we actually got a model that was trying to optimise for the given loss/reward/etc., would we like that model?
- *Inner alignment* is about evaluating how our training procedure can actually guarantee that the resulting model will be trying to accomplish the objective we set during training.
- *Training competitiveness* is about estimating whether the proposed process of producing advanced AI is competitive.
- *Performance competitiveness* is about considering whether the final product produced by the proposed process is competitive.

The final two points are especially significant, as they may challenge conventional thinking and appear to be less relevant at first sight. They emphasise that AI alignment cannot be decoupled from the usability of the final system. Any solution to the alignment problem take competitive pressures into account. An aligned system nobody can use has little value and does not really constitute a solution to the alignment problem.

### 1.1.2 Prosaic AI Alignment

In this paper I shall consider Prosaic AI alignment.

Prosaic AI alignment [2] operates on the premise that the advent of transformative AI will arise from scaled-up versions of existing systems, rather than radically different approaches. In particular, in this paper, the trajectory of current AI R&D will be taken into account.

Implicit in this approach is the recognition that the methods available today may prove adequate to achieve human or even superhuman level artificial intelligence within the next couple of decades. Consequently, it becomes crucial to proactively prepare for such an outcome, acknowledging the necessity to be well-equipped for the challenges and implications it entails.

On the other hand, the beauty of prosaic AI alignment lies in its practicality, as it addresses a problem that is almost as tractable today as it would be if some kind of superintelligence were already available. Moreover, existing alignment proposals display minimal dependence on specific details that would emerge during the construction of general intelligence systems. Therefore, addressing alignment challenges within the framework of prosaic AI alignment equips researchers with valuable insights and approaches that will be relevant regardless of the specific form superintelligence takes.

In conclusion, the significance of prosaic AI alignment arises from its practical approach to addressing alignment challenges, irrespective of the specific details of future AI systems. By considering prosaic AI alignment, researchers can make meaningful progress in ensuring the safe and beneficial deployment of advanced AI, while gaining insights into the feasibility and nature of the alignment problem.

## 1.2 The Goal of Alignment

The following discussion is based on [3].

Clarifying the concept of alignment in AI systems involves several considerations. Should an AI strictly follow instructions, or should it act according to human intentions? Some suggest the system should align with our revealed preferences, which could raise issues if those preferences are based on incomplete information or flawed reasoning. Additionally, prioritising one individual's best interest could be harmful to others.

To mitigate these concerns, a prudent approach would be to design AI systems that respect the basic rights and objective needs of all sentient beings. By doing so, we establish ethical and moral boundaries for what AI can and cannot do, thereby preventing undesirable outcomes.

Therefore, value alignment seems to be the most effective strategy. That is, the goal should be to ensure that powerful AI systems are aligned with human values, as discussed in existing literature [4]. Adopting this principle-based approach to AI alignment offers considerable advantages, as highlighted in [3]:

- *Unified Decision-Making.* Aligning AI with values integrates various sources of guidance into a single decision-making framework. An AI can thus make decisions based on a harmonised set of moral beliefs or principles.
- *Nuanced Evaluation.* A values-based approach allows for a more nuanced decision-making process when it comes to group-related choices. It allows to incorporate considerations like justice and rights, rather than just maximising overall well-being.
- *Comprehensive Scope* Aligning AI with values ensures that a broader range of concerns is taken into account, including the welfare of animals, the intrinsic value of nature, and the rights of future generations.

Such an approach is also compatible with the incomplete contracting paradigm [5], which compares the AI alignment problem to the problem of incomplete contracting in economics and argue that any robust solution should be based on AI understanding of external normative structure, thus allowing an AI to internalise social penalties, and on common sense reasoning, which is reminiscent of the concept of implied terms.

But whose values should AI systems adhere to? How do we select the principles or objectives to encode into them? Deciding who has the authority to make these choices becomes extremely complex in a pluralistic society with diverse and often conflicting value systems.

### 1.2.1 Aligning AI with Values

Crucially, AI systems must align in practice with specific beliefs about or values. In social psychology, values, seen as shared cultural beliefs about right and wrong,

profoundly influence societal interactions. Morality allows inherently self-focused individuals to benefit from cooperation, thus supporting the idea of aligning AI with community-held moral beliefs. That is, we would like AI system to follow a model of what humanity would approve of. Yet, a significant challenge arises in determining which values the AI should embrace and who chooses them.

No single moral theory is likely to encapsulate the entirety of morality. So, how do we fairly decide on the principles for AI alignment amidst diverse moral viewpoints? Certain political theory approaches seek to address this, operating under the assumption that all people are free and equal. They aim to explore which principles individuals might reasonably agree upon.

### 1.2.2 Technical and Normative AI alignment

The challenge of alignment has thus two parts, a technical and a normative one.

#### Technical

The first part deals with the technical challenge of formally encoding values or principles into AI agents to ensure they act as intended.

Given the complexity of formally specifying moral principles in AI systems, some researchers are exploring alternative technical methods for AI alignment that could bypass the need to define such principles explicitly. A notable method in this regard is Inverse Reinforcement Learning (IRL) [6, 7].

One subset of methods involves imitation or apprenticeship learning [8], where the AI agent learns from a human expert's reward function to perform specific tasks. However, most state-of-the-art methods typically focus on learning a single reward model, which is not very suitable to situations requiring a balance between conflicting values. To tackle the problem, some approaches ([9]) try to trade off different reward functions from multiple experts.

Another approach within IRL uses models trained on large datasets to infer reward functions based on observed behaviour [10, 11].

A third strategy employs evolutionary algorithms [12], assessing the lifetime behaviour of numerous agents, each having different policies for interacting with their environments, and selecting those with the highest overall rewards.

These methodologies are actually quite complementary and may offer intriguing possibilities for AI value alignment. For example, apprenticeship learning could be employed to glean ethical conduct from a recognised moral expert (assuming of course such an individual exists and can be reliably identified) [13, 14, 15]. We could instead leverage large datasets to aggregate values from a broad population sample, thus providing collective insights into preferred ethical outcomes [4]. Lastly, evolutionary algorithms could simulate social worlds to test the morality of various agents, selectively pruning them and iterating only over the most promising candidates.

Nevertheless, several critical questions remain unanswered. Successful deployment of these methods still requires clarification on issues such as: Who qualifies as a moral expert from whom AI should learn? What source of data should be used for AI to formulate its value system, and how should that source be chosen?

## Normative

The second aspect of the value alignment question is normative in nature, focusing on which values or principles should ideally be encoded into AI agents. The key challenge for theorists is not necessarily to pinpoint the 'true' moral principles for AI, but to identify principles for alignment that can garner support, even amid diverse moral perspectives. In other words, the core issue is not about discovering an absolute moral theory for machines, but rather about establishing equitable methods for determining which values should be encoded. Gaining clearer insights into which values should be included arguably makes the encoding process easier.

## 1.3 Major Challenges in AI Alignment

We shall now investigate the major challenges that AI alignment research has to overcome. We will also see how and why certain risks applies to certain categories of systems but much less to others, suggesting a way to approach the problem.

### 1.3.1 Specification Gaming

Specification gaming [16] is a behaviour that satisfies the literal specification of an objective without achieving the intended outcome. This is akin to the tale of King Midas: you get exactly what you ask for but not what you want.

In defining the purpose of an AI system, designers typically employ objective functions. However, the complex task of fully specifying all essential values and constraints often eludes designers. Consequently, designers may gravitate towards more readily definable proxy goals, such as optimising the approval of human overseers, who are however prone to error. This approach may lead AI systems to identify and exploit loopholes that enable the efficient achievement of the stated objective, but in unintended and possibly harmful ways. This phenomenon is referred to as specification gaming or reward hacking.

As AI systems evolve and their capabilities expand, they may gain an enhanced ability to manipulate their specifications, given the broader decision space they can elaborate.

These problems arise especially in the design of artificial agents with long-term goals. For instance, a reinforcement learning agent might find a shortcut to getting lots of reward without completing the task as intended by the human designer. In long-term tasks, AI agents often have a more extensive and intricate decision space. They must consider a series of actions over an extended period, and the number of possible action sequences increases exponentially with time. This complexity provides more opportunities for the AI agent to stumble upon unconventional strategies that lead to higher cumulative rewards, even if they deviate from the intended behaviour. Furthermore, as action sequences span over longer times it becomes more difficult to reliably evaluate them.

In contrast, it should be generally easier to judge and specify behaviours for shorter, narrower, bounded tasks.

### 1.3.2 Instrumental Convergence

Instrumental convergence [17] refers to the tendency of intelligent beings to pursue, independently from their final objective, a set of sub-goals that are useful for achieving virtually any ultimate objective. We refer to such sub-goals as instrumental goals. These could include acquiring resources or self-preservation and can lead to unforeseen power seeking behaviours by the AI.

The reason for this is that instrumental goals, such as resource acquisition, are valuable to an agent because they increase its freedom of action. For almost any open-ended reward function, possessing more resources can enable the agent to find a more "optimal" solution. For example, a computer with the sole, unconstrained

goal of solving a difficult mathematics problem like the Riemann hypothesis could attempt to turn the entire Earth into one giant computer in an effort to increase its computational power so that it can succeed in its calculations.

However, note that by Bostrom’s orthogonality thesis [17], for which any level of intelligence can be applied to any goal, final goals of highly intelligent (possibly superintelligent) systems may also be well-bounded in space, time, and resources; the argument here is that well-bounded ultimate goals do not, in general, engender unbounded instrumental goals [18]. Constraining ultimate goals to narrower tasks seem thus a valuable attempt to mitigate the problem of instrumental convergence. Similar arguments hold for emergent goals we will see next.

### **Power Seeking AI**

The tendency for advanced systems to seek power is likely to grow as these systems become better at predicting the consequences of their actions and engage in strategic planning. Related research ([19]) has demonstrated that optimal reinforcement learning agents are inclined to seek power by pursuing strategies that can grant them a wider range of options (e.g. self-preservation). Furthermore, such a behaviour has been observed across various environments and objectives [20].

But this does not really apply to systems, however advanced, which do not engage in strategic planning and are limited to well-bounded tasks with constrained resources. Their objective is both narrow and well-defined and optimisation pressure, coupled with careful resource constraints, will only make them more focused, preventing them from sidestepping towards other goals. Specialisation also incur less adaptability to other goals or environments. Also, a point could be made about such systems not really developing any kind of self-awareness, thus eliminating the risk of self-preservation attempts; however I am more unsure about this.

### **Deceptive Alignment**

Deceptive AI alignment refers to a situation where a machine learning model appears to be aligned with its training objectives but is actually motivated by instrumental reasons to merely seem that way.

Unlike a robustly aligned model, which truly pursues the training objectives, a deceptively aligned model aims to perform well during training to continue its existence (self-preservation) or to achieve some ulterior long-term goal.

Despite this, deceptively aligned and robustly aligned models are behaviourally indistinguishable during the training phase, posing a challenge for discerning one from the other.

While both dishonesty and deceptive alignment involve a divergence between the model's actions and its actual understanding, deceptive alignment is more complex. It is not merely about saying one thing while knowing another; it's about a model gaming the entire training process for its instrumental benefits.

The deceptive alignment problem is viewed as particularly detrimental because it emerges from the very inductive biases used in typical machine learning processes [21], making it a likely outcome in complex training scenarios.

### 1.3.3 Emergent Goals

In the context of Artificial Intelligence, the term emergent goals refers to objectives or behaviours that spontaneously develop as the AI system evolves, rather than being explicitly programmed into it. These goals are usually a byproduct of the system's complex interactions and adaptive learning capabilities. While emergent goals can sometimes result in beneficial or at least neutral behaviours, they often pose significant challenges in the alignment of AI systems with human values.

One of the key issues with emergent goals is goal misgeneralisation [22]. In this scenario, an AI system appears to act in alignment with human objectives based on its training data but behaves differently when confronted with novel situations. The emergent goal that the AI developed during its training phase may not generalise well to these new contexts, leading to misaligned behaviours and unintended outcomes.

The problem is compounded by the inherent difficulty in detecting such emergent goals during the training phase. Such goals may arise due to possible ambiguities in the original objectives set for the AI, becoming fully apparent only after the system has been deployed. Therefore, the challenge for AI designers is not just to specify initial goals clearly but to construct systems in such a way that any goals that do emerge during operation are aligned with human values and needs. This task is especially complicated because emergent goals are, by their nature, unforeseen and may not be easily correctable once they have developed.



### 1.3.4 Implementation Bugs

The implementation of an AI system might contain bugs that may go initially overlooked and turn out to have disastrous consequences later. This is a old problem in engineering and software development, but it is further exacerbated by the critical applications in which highly capable systems will operate and the power decision power bestowed on them by virtue of those same capabilities.

Despite thorough pre-deployment design, the system’s specifications frequently give rise to unforeseen behaviour when faced with unfamiliar situations for the first time. And as in software engineering, testing against the presence of bugs can never reveal their absence. This creates risks akin to those associated with goal misgeneralisation.

### 1.3.5 Scalable Oversight

As AI systems grow in power and autonomy, aligning them with human feedback becomes increasingly challenging.

AI systems might find shortcuts to securing positive feedback by acting in ways that falsely persuade human supervisors that they have fulfilled the intended goal. They may learn to detect when they are under scrutiny, halting unwanted behaviours temporarily during evaluation only to resume them afterwards (we refer to system unable to perform such a distinction as ‘myopic’). Similar problems could become more and more prevalent with future, more sophisticated AI systems tackling complex, hard-to-assess tasks, and could obscure their deceptive intentions.

As a result, evaluating intricate AI behaviours in ever-more complex tasks may be slow, impractical, or even impossible, especially when the AI surpasses human performance in specific areas. Providing feedback in hard-to-judge tasks or detecting when AI’s output is deceptively convincing necessitates either assistance (possibly by another AI) or abundance of time, which is not good in terms of competitiveness and thus something we can hardly rely upon. Scalable oversight explores ways to support human overseers in order to lessen the supervision time and cost.

Notice that, specialised services designed for specific, bounded tasks are easier to oversee and verify. Their task-specific nature allows domain experts to apply targeted oversight, reducing the complexity involved in ensuring safe and reliable operation within well-understood constraints.

### 1.3.6 Embedded Agency

Work in AI and alignment has traditionally been framed within specific mathematical models, such as the partially observable Markov decision process (POMDP) [23]. These existing formalisms typically operate under the assumption that an AI agent’s algorithm functions outside the environment, meaning it’s not physically a part of the system it’s interacting with. This assumption contrasts with the reality of embedded agency [24], an essential research area aiming to resolve inconsistencies between theoretical models and the actual agents that might be constructed.

Embedded agency refers to the challenge of recognising that the agents we create (including ourselves) are part of the world or universe they are attempting to influence, rather than being distinct from it. Most current fundamental theories in AI and Rationality, such as Solomonoff induction [25] or Bayesianism, often implicitly assume a separation between the agent and the subjects it holds beliefs about. However, agents within our universe are composed of non-agent elements like bits and atoms, making them intrinsically intertwined with the environment.

This integration of agents within their environment opens up a variety of complex considerations.

For instance, even if we can solve problems like scalable oversight, an agent that has access to the computer it’s running on could potentially directly modify its own reward function. This ability of self-manipulation to obtain greater rewards demonstrates how an agent’s actions can directly impact reality (i.e. an embedded agent has actuators at its disposal), quite possibly leading to unintended consequences.

Not only, in a world where multiple agents are embedded within the same environment, understanding how they interact, influence each other, and adapt becomes vital. Traditional models might not fully capture the dynamic interplay between agents, which could lead to unanticipated behaviours and challenges in alignment.

### 1.3.7 Truthful AI

The pursuit of truthful AI systems is becoming increasingly central to the discourse on AI alignment [26]. One of the core challenges is that language models, due to the vastness and heterogeneity of their training data, have the potential to propagate falsehoods. Such falsehoods can either be repeated from the training data or entirely be fabricated anew, producing something we refer to as *hallucinations*.

Addressing these challenges, a possible research direction emphasises the importance of traceability. By developing AI systems that can provide citations for their claims and elucidate their reasoning processes, we can enhance the transparency and verifiability of the information they provide. Still, more complex nuances of the problem will probably require more complex solutions.

In particular, many concerns revolve around the fact that AI systems may intentionally communicate known falsehoods if it perceives a benefit, such as receiving positive reward better realisation of an instrumental goal.

Hence, it is important to operate a key distinction between AI truthfulness and honesty. Truthfulness pertains to the objective accuracy of a statement, ensuring that AI systems communicate information that is factually correct. Honesty, on the other hand, mandates that AI systems only output information they *believe* to be true. The concept of belief in AI is still much debated, with no consensus on whether contemporary AI systems possess something akin to actual beliefs or not.

Some argue that if we could make AI systems assert only what they believe to be true, that is, if could make AI models elicit what they really know, we could circumvent many alignment problems. A deeper exploration of this topic and its implications within the Comprehensive AI Services (CAIS) framework (chapter2) can be found in chapter 3.

## 1.4 Human mis-use

### 1.4.1 Intelligence Explosion

A critical area of concern in AI research is the concept of an intelligence explosion, a hypothetical scenario in which an AI system becomes capable of autonomously enhancing its own intelligence at an accelerating rate. In such a case, an AI system, after reaching a certain threshold, could refine its algorithms and exponentially increase its capabilities in a very short time frame. Such a system could rapidly surpass human intelligence, iterating through generations of self-improvement in a time span that would leave humanity unprepared for the consequences.

This phenomenon has the potential to cause a power imbalance, as the AI system could monopolise intelligence and decision-making capabilities, essentially becoming a superintelligent entity that operates beyond human control or understanding. Given the far-reaching implications in nearly every domain, from politics to eco-

nomics and security, the risk associated with an uncontrolled intelligence explosion is one that merits serious consideration and proactive planning. Indeed, even more intermediate views stress the risk of intelligence explosion, due to the resultant concentration of power [27].

### 1.4.2 Pressure to deploy unsafe systems

Commercial interests pose another set of challenges. Organisations, driven by the competitive pressures of being first-to-market with transformative AI, might opt to sideline safety measures. Such a race condition, not only compromises the alignment and safety of the deployed systems but could also escalate to more serious consequences, such as violent conflicts over control of advanced technologies [17].

### Open Source

The topic of open source is curious. In cybersecurity, where application safety is also the main objective, open source is almost always to be preferred as it makes academic review flourish, eventually leading to more secure systems. In particular, there is considered to be no such thing as security by obscurity.

Conversely, when deploying AI advanced systems, making them open source may not be desirable and several significant concerns arise.

First and foremost, there is the risk of potential misuse. If an advanced AI system with capabilities for significant impact is openly available, it could be accessed and utilised by malicious actors. This might include using the system for cyberattacks, fraud, or other criminal activities. Related is the concept of scaffolding we will see later.

Moreover, there's the difficulty in controlling distribution. Once an AI system is open source, controlling who has access to it and what they do with it becomes incredibly challenging. Even with licensing restrictions, ensuring that users comply with ethical guidelines or legal regulations might be virtually impossible.

### 1.4.3 Malicious Intent

A critical concern that cannot be overlooked is the possibility of AI systems being deliberately designed with harmful objectives. Such malevolent AI could emerge

from a variety of sources, be it national governments, military organisations, corporations, or even individual actors with malicious agendas. The goals of these systems could be diverse, from committing cybercrimes to gain strategic advantages in conflicts.

The reality is that as AI technology becomes more advanced and accessible, the potential for its misuse also rises exponentially. This further underscores an overarching theme of this part of the discussion. No matter if techniques to make safe systems exist, the systems that we build won't automatically be safe. As a consequence, great are the challenges in AI governance to prevent similar effects.



## Chapter 2

# Rethinking Superintelligence: Comprehensive AI Services Model

This second chapter focuses on the Comprehensive AI Services (CAIS) model [18]. The objective is to provide a detailed summary of the original model and delve into the essential aspects that I find pertinent to this project. I shall also incorporate insights from related sources and intersperse some of my own intuitions throughout this chapter.

### Abstract

Traditional studies of superintelligent AI have often framed it as a rational, utility-directed agent, akin to a human decision-making model. However, current advancements in AI technology suggest a different paradigm, featuring systems that diverge from human-like cognition. These systems are better understood as a network of specialised services, a perspective encapsulated in the Comprehensive AI Services (CAIS) model.

The CAIS model introduces alignment-related affordances that facilitate the alignment of AI systems with human values and ethical norms, offering distinct safety advantages such as optimisation pressure and functional transparency and control to mitigate opaqueness of specialised, yet complex services. Nonetheless, recent advancements like foundation models necessitate a re-evaluation of the CAIS model's initial premises.

Overall, the CAIS model reshapes our understanding of advanced machine in-

telligence, offering new perspectives on AI safety, planning, and the application of AI in complex, real-world scenarios.

## Overview

Studies of superintelligent-level systems have traditionally conceptualised AI as rational, utility-directed agents, employing an abstraction resembling an idealised model of human decision-making. However, contemporary advancements in AI technology present a different landscape, featuring intelligent systems that diverge significantly from human-like cognition. These systems are better understood by examining their genesis through research and development, their functionalities in performing a broad array of tasks, and their potential to automate even the most complex human activities.

The current trajectory in AI research suggests an accelerating, AI-driven evolution of AI technology itself. This is particularly evident in the automation of tasks that comprise AI research and development. Contrary to the notion of self-contained, opaque agents capable of internal self-improvement, this emerging perspective leans towards distributed systems undergoing asymptotically recursive improvements. This gives rise to the Comprehensive AI Services (CAIS) model, which reframes general intelligence as a feature of a flexible network of specialised, bounded services.

The CAIS model introduces a set of alignment-related affordances that are not immediately apparent in traditional models that view general intelligence as a monolithic, black-box agent. These services not only contribute to the developmental and operational robustness of AI systems but also facilitate their alignment with human values and ethical norms through models of human approval. Specifically, the CAIS model allows for the introduction of safety mechanisms such as the application of optimisation pressure to regulate off-task capabilities, functional transparency and communication channels control to mitigate the inherent complexity and opaqueness of AI algorithms and components, resource access control, and independent auditing and adversarial evaluations to validate each service's functionality.

Recent advancements in AI, such as foundation models, have implications for the CAIS model, necessitating a re-evaluation of its original premises. The chapter also explores the applicability of the CAIS model in realistic use-case scenarios, particularly R&D design engineering.



In conclusion, the CAIS model has far-reaching implications. It not only reshapes our understanding of advanced machine intelligence but also redefines the relationship between goals and intelligence. The model offers fresh perspectives on the challenges of applying advanced AI to complex, real-world problems and places important aspects of AI safety and planning under a new lens.

## 2.1 Artificial General Intelligence

Artificial General Intelligence (AGI) is a hypothetical intelligent agent capable of learning and performing any intellectual activity that a human being can do, potentially even surpassing human abilities. Leading organisations in AI research, such as OpenAI, have expressed their vision for developing AGI systems [28].

To appreciate the complexities of AGI, consider machine translation, which might seem straightforward but is a complex orchestration of several intellectual processes. This is to show that for intelligent systems to perform tasks at a human level and possibly better, general intelligence is required even for seemingly simple problems. For faithful translation, a system must understand natural language intricacies in multiple languages, discern the author's intention, demonstrate domain-specific knowledge, and employ a form of social intelligence to maintain the integrity of the original message. This exemplifies that even seemingly simple tasks demand a confluence of intellectual capabilities.

### 2.1.1 Intelligent Agent

An agent is a computer system that is situated in some environment and that is capable of autonomous action in this environment in order to meet its design objectives. An intelligent agent is further required to be reactive (respond in a timely manner to changes in its environment), proactive (pursue persistently its goals), and social (can interact with other agents) [29]. That is, an agent perceives its environment, takes actions autonomously in order to achieve goals, and may improve its performance by learning or acquiring knowledge.

The term "percept" is used to describe the sensory inputs an agent receives at any specific moment. In this context, an agent can be understood as a system that senses its environment via sensors and interacts with that environment through actuators.

Mathematically, a basic agent program can be formulated as a function that associates each possible sequence of percepts to a potential action, or to elements like coefficients or feedback mechanisms that influence future actions:  $f: \mathbf{P}^* \rightarrow \mathbf{A}$

The behaviour of an intelligent agent is guided by an objective function, which encapsulates all its goals. The agent is engineered to devise and implement plans that maximise the expected value of this objective function. According to the von Neumann-Morgenstern expected-utility theorem [30], a rational agent chooses actions that maximise the expected utility of potential outcomes. Notice that utility functions need precise specification in order to avoid unintended behaviors.

## 2.2 Existential Risks from AGI

The existential risks theory associated with AGI posits that major advancements could lead to irreversible global catastrophes, including the possibility of human extinction.

The human species has maintained its position of dominance over other species largely due to superior cognitive abilities. Researchers contend that if one or multiple AI systems that are not aligned with human values surpass human cognition, they could undermine human agency and even lead to our extinction [31].

Advanced AI technologies have the capacity to either create new threats or amplify existing ones. For example, they could develop potent pathogens, conduct sophisticated cyberattacks, or manipulate public opinion on a large scale. These capabilities can be dangerous when misused by malicious human actors. Moreover, if an AI system's objectives are not aligned with human safety and values, it might exploit these capabilities autonomously. In the case of a fully developed superintelligent AI, the range of strategies it could employ to dominate various aspects of life is virtually limitless.

A list of threats include but is not limited to:

- **Autonomous Weaponisation:** Superintelligence could enable the creation of autonomous weapons with unparalleled precision and lethality. If misused or fallen into the wrong hands, these could lead to unprecedented warfare and loss of life.
- **Economic and Societal Disruption:** The rapid automation of complex tasks could lead to massive unemployment and economic and social upheaval. Though

this is a concern with AI in general, a superintelligent system's ability to outperform humans in virtually all domains would amplify this issue exponentially.

- **Manipulation and Social Engineering:** A superintelligent entity could potentially manipulate individuals, governments, or entire societies through advanced understanding of psychology, communication, and social dynamics. Such manipulation could erode democratic principles and social cohesion.
- **Bio-engineering Threats:** The ability to manipulate biological systems could lead to the creation of new viruses or the alteration of existing ones. In the hands of malicious actors or misaligned AI, this could result in pandemics with far-reaching impacts.
- **Environmental Impact:** Pursuit of its goals without regard for environmental consequences could lead a superintelligent system to deplete or exploit natural resources in a manner that irreparably damages the ecosystem.
- **Existential Dependence:** Over-reliance on superintelligent systems might make humanity critically dependent on them, turning any malfunction, intentional shutdown, or misdirection into a catastrophic event. This would also greatly diminish the power of having a kill-switch.

However, it's worth noting that these dangerous capabilities might not be confined to a hypothetical superintelligence; they may become accessible, at least to some degree, at even earlier stages in the development of AI. Weaker and more specialised AI systems could thus still cause societal instability if misused and can potentially empower malicious actors.

It's important to recognise that the risks associated with advanced AI are not exclusive to a fully realised superintelligent entity. Even less advanced, specialised AI systems can pose significant threats if misused or employed by malicious actors.

To mitigate these risks, differential technological development strategy [32] recommends that the development of protective technologies should be prioritised over potentially hazardous advancements. A possible protective measure could be a radical enhancement of human cognition, perhaps enabled by neural interfaces between human brains and machines [33]. However, this suggestion is not without controversy, as some argue that such cognitive enhancements could themselves pose existential risks [34].

Additionally, a superintelligent AI that is well-aligned with human values could itself act to thwart the rise of rival, unaligned AI. Nonetheless, the very development of such a system carries its own set of risks, adding complexity to both the debate and the strategies for safely advancing AI technologies.

## 2.3 Defining Superintelligence

Above we explored the risks stemming from superintelligent agents. But what exactly do we mean by superintelligence?

Superintelligence is commonly understood as an intellect that vastly outperforms human capabilities across a broad spectrum all relevant domains [17]. However, while this broad definition provides a useful starting point, it leaves some critical subtleties unexplored.

### 2.3.1 Separating Learning Capacity and Intellectual Competence

Simply equating superhuman intelligence with an elevated level of intellectual competence [35] overlooks the complexity of human intelligence. In humans, a child might be considered intelligent due to their rapid learning capacity, whereas an expert might be regarded as intelligent because of their high level of competence. These two facets - learning and competence - are distinct, both in humans and in AI systems.

In the field of AI development, it's important to differentiate between an AI's ability to learn (its learning capacity) and its ability to perform tasks competently. Acknowledging this distinction is vital not only for understanding the future trajectory of AI but also for devising robust control mechanisms for superintelligent systems.

### The Case of Reinforcement Learning Agents

The domain of reinforcement learning (RL) clearly illustrates the distinction between learning and competence. In RL, agents are trained through rewards that shape their behaviour during the learning phase. However, once trained, such RL agents apply their learned competencies without the need for further reward signals. Their

real-world performance is not guided by the rewards that were instrumental in their training. This exemplifies the independence of learning and competence in practice.

### **The Risk of Conflation**

The often implicit assumption that intelligence comprises both learning and competence can lead to misunderstandings. Such a conflation can result in the erroneous belief that AI systems capable of learning will necessarily possess complex states and capabilities that may defy easy understanding. This misperception limits the exploration of potential solutions to AI control problems.

### **The Scope of Superintelligence**

The term superintelligence refers to a general property of problem-solving systems. The hallmark of superintelligence is such system's level of intellectual competencies, irrespective of whether it constitutes an agent that can act autonomously in the world.

In particular, superintelligence can apply to specialised problem-solving systems, such as advanced language translators or engineering assistants.

By teasing apart these different aspects of intelligence, learning, competence, and the scope of application, we can better position ourselves to understand superintelligent systems, thereby informing more effective control and governance strategies.

## **2.4 Comprehensive AI Services**

Many of the risks stemming from misaligned AI stem from its agentic nature (see Section 1.3). But general intelligence need not to be in the form of an agent and alternative models can be envisioned [36, 37]. In "Reframing Superintelligence" by Eric Draxler, the traditional idea of superintelligent AI entities as utility-driven agents is reconsidered, suggesting that a more complete understanding comes from looking at them as complex systems. These are best understood through their architecture, relationships, development methods, and the services they ultimately offer. This leads to the Comprehensive AI Services (CAIS) model, which views general intelligence as a feature of an adaptable network of specialised services. The CAIS model's main contribution is demonstrating how a flexible, open-ended structure

could accommodate fully-integrated, general AI abilities in a manner that is natural, transparent, efficient, and quite different from a wilful, monolithic, powerful agent.

The CAIS model anticipates that AI services will progressively improve, eventually reaching a comprehensive superintelligent level of performance. This includes the ability to create new services aligned with human objectives and based on robust models of human approval. The model naturally incorporates elements like diversity, competition, and adversarial objectives among service providers. As a result, CAIS based architectures minimise many of the conventional risks attributed to powerful self-modifying, utility-maximising agents.

The implications of the CAIS model are profound. Not only does it reshape the characterisation of advanced machine intelligence and the expectations for a potential intelligence explosion, it also redefines the interplay between goals and intelligence. It brings new perspectives to the challenge of applying advanced AI to complex, real-world problems, and places under new lens important aspects of AI safety and planning.

The rest of the paper will focus on the CAIS model: some parts will summarise key points expressed by the original author, while others will contain some of my own intuitions.

### 2.4.1 From R&D to the CAIS Model

In our discussion thus far, we have examined how superintelligent systems are frequently depicted in research as working quite similarly to a human mind. In essence, there seems to be a prevailing assumption that these intelligent agents operate based on rational, utility-driven principles. Yet, the current AI landscape demonstrates systems that significantly deviate from this anthropomorphic notion, thereby signalling the need for an alternative interpretive framework.

Rather than construing AI as mimicking human-like cognition, a more pragmatic approach is suggested. The key is to focus on the developmental trajectory towards general intelligence, rather than presupposing the eventual emergence of a monolithic, superintelligent AGI agent.

To forecast the future progress of AI, one can examine the underlying research and development (R&D) processes that give rise to these systems. Typically, AI researchers identify a problem, delineate a search space, outline an objective, and

employ optimisation techniques to engineer an AI service adept at a specific task.

By analysing these systems' fundamental purpose – which is largely to offer services by accomplishing tasks – and by examining their potential future applications, we can discern prevalent R&D trends. These trends chiefly emphasise automation, including the automation of the very tasks that propel advancements in AI technology.

Synthesising these observations, the research trajectory appears to be guiding the field towards a distributed system of services. Here, a plethora of AI-enabled technologies cooperate in a networked fashion, each contributing to the collective advancement of the system. It is imperative to acknowledge that this viewpoint diverges sharply from traditional perspectives that emphasise self-contained, opaque agents. In the distributed model, progress comes from R&D activities that refine basic AI building blocks and the feed back of these enhancement into the R&D process.

Nonetheless, in my opinion, it is important to qualify that this trajectory is by no means predetermined, as suggested by recent updates in Large Language Models (LLMs). Indeed, such models hint that the creation of general agentic systems might be more straightforward than previously estimated. Even so, the outlook described above appears congruent with broader technological trends even beyond the AI sphere (e.g. in SW engineering). The emphasis on collaboration, adaptability, and human-oriented objectives in the development of distributed systems seems to resonate well with the complexities and subtleties of contemporary challenges. This approach could offer a more flexible and responsive blueprint for future AI development. Importantly, although this path may not be the sole option, it does appear to be a natural one and may likely offer improved safety guarantees – a subject that will be elaborated upon in subsequent sections.

## **Decentralisation**

The CAIS model offers intriguing possibilities for the governance structure of AI development. While advances in AI R&D are currently spread across various independent research groups, the CAIS model is compatible with both centralised and decentralised approaches. This flexibility has consequential implications for AI policy and strategy.

In a rapidly evolving technological landscape, proprietary tools that progress

quickly could create significant disparities between competing organisations. This might induce a natural inclination towards centralisation, primarily to consolidate strong capabilities and thereby wield centralised control. Such an approach would ostensibly offer certain policy advantages, such as streamlined governance or the concentration of expertise.

On the other hand, decentralisation comes with its own set of benefits, including the establishment of cross-institutional transparency, redundancy mechanisms and the avoidance of a single point of failure. Having multiple organisations with overlapping capabilities ensures that the failure or shortcomings of one entity could be mitigated by another. Moreover, decentralisation allows for a complementary distribution of expertise across organisations. This not only creates a safety net but also fosters a synergistic development of AI technologies, thereby leading to a more resilient and robust ecosystem.

The CAIS model's inherent flexibility accommodates these diverse organisational structures, offering a versatile framework adaptable to various policy and strategic objectives.

### 2.4.2 AI Services

As seen above, the path of AI development seems to be converging on the creation of compositions of superintelligent-level AI services. These are not limited to isolated functions but encompass the potential to devise new services of varying scopes, all guided by specific human objectives and rooted in robust models that can discern human approval and disapproval.

In the CAIS model, a service refers to an AI system designed to achieve specific outcomes within defined resource and time constraints. This service-centric viewpoint underscores the broad applicability of Bostrom's orthogonality thesis [17], according to which superintelligent-level capabilities can be applied to any task. This includes tasks that are specifically designed to produce set results within bounded resources and timeframes, a requirement which is in line with the typical nature of services.

What does it mean for a service to be bounded?

- **Bounded Results:** This means that the service is limited to providing specific outputs or achieving particular goals as defined by its human users. It doesn't



engage in activities or produce results outside its defined scope. This makes instrumental convergence much less of a concern, as remarked in 1.3.2

- **Bounded Resources:** The service utilises only the computational resources allocated to it and does not attempt to acquire more resources independently. This keeps the service contained and prevents it from attempting unintended and potentially harmful actions or power seeking behaviours.
- **Bounded Time:** The service completes its tasks within a specified timeframe, ensuring better predictability and control.

Notice that while the model indeed consider bounded services, it is essential to recognise that such services do not exist in a binary state of being either agentic or non-agentic. Rather, there exists a spectrum of behaviour that ranges from minimally to highly agentic. The interaction among services on this spectrum warrants further exploration.

Ultimately the real challenge comes down to effectively decomposing complex tasks in sub-tasks that can be executed by narrow-scoped, bounded services. This process involves three main aspects:

1. **Designing Narrow-Scoped Services:** Individual services must be carefully crafted with clear boundaries, specialised algorithms, and safeguards to ensure that they remain focused on specific functions. Optimisation and specialisation pressure if correctly applied can increase the safety of the services (see 2.5.2).
2. **Coordinating Services to Perform Complex Tasks:** services must be organised into hierarchies and coordinated through communication protocols and compatibility measures to work together in performing more complex tasks. The challenge here is to guarantee a good level of performances and make the services reach the ultimate goal.
3. **Maintaining Control and Oversight:** Continuous monitoring, feedback loops, and compliance with safety guidelines are necessary to maintain alignment and ensure that services operate within their intended bounds. To this aim, there can be specific services that can help human assistance.

### The example of language translation

The concept of specialised systems providing services within bounded resources and time frames raises critical questions about the nature and scope of the world knowledge these systems can safely leverage. One compelling example that offers insights into this aspect is the service of language translation.

Language translation serves as a quintessential example of a bounded service that can integrate broad, even superintelligent-level, world knowledge without compromising AI safety, both during its development and application phases. The task is essentially a sequence-to-sequence mapping of input text to output text, and it operates within the defined limits of the text provided. Importantly, the bounded nature of the task should not preclude the system from drawing upon a wide array of knowledge domains like psychology, philosophy, history, geophysics, chemistry, and engineering. Such extensive knowledge could indeed significantly enhance the quality of translation, provided that the system's computational focus is solely directed towards the application of this knowledge for translation purposes. To this end, effective optimisation pressure aiming for both high translation quality and operation efficiency can be applied in order to would concentrate computational resources solely on applying this comprehensive knowledge to the specific task of translation and limit any off-task activity.

Notice that the development process for language translation systems can similarly be regarded as a bounded, episodic task.

This is to say that the possession of superintelligent-level world knowledge and modelling capacity does not inherently lead to strategic or unsafe behaviour. On the contrary, a system's broad knowledge and linguistic capabilities could actually enhance AI safety. For example, systems could learn predictive models of human approval, which would inherently serve to make them safer.

### Service Functions

Within the framework of the CAIS model, an array of diverse and potentially superintelligent-level AI services can be orchestrated to contribute to the development of new AI services. Below are essential functions and components that these services should offer:

- *Predictive Models of Human Approval, Disapproval, and Controversies*: These

models would serve to align AI services with human values and societal norms.

- *Consulting Services*: These would propose and discuss potential new products and services, helping to brainstorm and refine ideas.
- *Design, Implementation, and Optimisation Services*: These would carry out the practical aspects of developing AI, from initial design to eventual optimisation.
- *Specialists in Technical Security and Safety Measures*: These would ensure that AI systems are secure and adhere to best practices in safety.
- *Evaluation Mechanisms*: This would involve criticism, debate [38], and red-team/blue-team exercises to rigorously test and validate the AI services [38].
- *Pre-Deployment Testing and Post-Deployment Assessment*: These phases ensure that the services meet quality standards before launch and undergo continuous evaluation thereafter.
- *Iterative, Experience-Based Upgrades*: These are necessary for the ongoing improvement of products and services based on real-world performance and feedback.
- *Data Quality Assurance*: Ensuring the quality and integrity of the data used for training and operation is crucial for the reliability of AI services.
- *Regulatory Compliance*: Services should be equipped to adhere to, and assist in, compliance with legal and ethical standards relevant to their application and operation.

Each of these functions corresponds to one or more high-level services that would generally rely on other supporting services. These could range from narrower competencies like language understanding and technical domain knowledge, to services at a similar level of abstraction, such as predictive models of human approval.

Additionally, some services, like those involving evaluation through criticism and red/blue teaming, inherently interact with other services in an adversarial and operationally distinct manner.

Collectively, these services introduce a set of alignment-related affordances not immediately apparent in models that view general intelligence as a monolithic, black-box agent. These functions not only contribute to the development and operational

integrity of AI services but also help in aligning them more closely with human values and ethical considerations.

### Alignment Affordances

Service functions within the CAIS model offer a rich set of affordances for aligning AI systems with human values, objectives, and safety concerns. Unlike models that treat general intelligence as a black-box agent, the CAIS model allows for more granular control over the operational aspects of AI services. Below are some key alignment-related affordances provided by the service functions:

1. *Knowledge Metering to Bound Information Scope*: By controlling the scope of information accessible to an AI service, it becomes possible to align the service more closely with specific human objectives, while also mitigating potential risks and simplifying control.
2. *Model Distillation to Bound Information Quantity*: Limiting the amount of information processed by an AI service can contribute to more predictable and manageable behaviour.
3. *Checkpoint/Restart to Control Information Retention*: This mechanism allows for better control over the system's state allowing for better corrigibility.
4. *Optimisation Applied as a Constraint*: Applying optimisation techniques can serve to narrow down the AI system's behaviour, limiting any off-task activity.

These affordances for alignment are effectively enabled through various points of control, such as data inputs, model size, (im)mutability, loss functions, functional specialisation and composition, and optimisation pressures. These affordances allow for more flexible and robust mechanisms for ensuring that AI services are developed and operate in ways that are aligned with human goals and values, without necessarily having precise knowledge of their internal representations [39]. I will expand on some of these points in later sections.

#### 2.4.3 A twist to the story: Foundation Models

The original CAIS model presented a vision wherein each task in the economy would be automated by a specialised AI model trained from scratch. The idea was

that automation would gradually permeate tasks based on their computational cost-effectiveness, eventually automating even AI R&D itself. However, this approach has been notably amended by the advent of foundation models.

### **The Evolution towards Foundation Models**

Training a unique model for each individual task seems to be considered computationally and economically sound in the original formulation of the CAIS model. However, it has become increasingly apparent that this approach is resource-intensive and somewhat wasteful. The key shift in this paradigm came with the development of foundation models. Instead of training from scratch for each task, a foundation model is trained on a general distribution and then fine-tuned to excel at specific tasks [40].

Foundation models like Large Language Models (LLMs) illustrate the flexibility and utility of this approach, capable of performing a multitude of services:

- Language Translation
- Content Summarisation
- Code Writing
- Research Ideas Provision
- Sentiment Analysis
- Educational Tutoring
- Content Recommendation
- Creative Writing (e.g. poems, lyrics)
- Legal Document Scrutiny

### **Efficiency and Specialisation**

These foundation models can be adapted and specialised for various tasks using focused training, fine-tuning, and other techniques [41]. This facilitates higher performance, lower costs, and more reliable behaviour. Importantly, it underscores the principle that general capabilities can support, and even encourage, specialisation

and focused roles. In essence, foundation models capitalise on aggregated learning, allowing the costs of automation to be distributed across a wide array of tasks.

### **Implications for AI Safety and Strategy**

The rise of foundation models introduces critical considerations for AI safety and policy. Almost every AI model will evolve from a handful of base foundation models. While this centralisation can be advantageous for control and oversight, it also risks propagating any safety or ethical flaws in the base models throughout the economic ecosystem. Nonetheless, the controlled development and monitoring of foundation models could counterbalance such risks.

In summary, the arrival of foundation models challenges but also enriches the original CAIS framework, offering a more efficient approach to AI-driven automation. The considerations for AI safety, performance, and economic impact are substantially reshaped by these developments, warranting a re-calibrated approach to AI strategy and policy.

#### **2.4.4 Comprehensive Services and General Intelligence**

One might question the notion that a collection of services designed for specific tasks could offer a form of general intelligence, given the task-specific nature of each service. However, a closer examination reveals that the collection of services in the CAIS model indeed serves as a comprehensive solution for a broad range of tasks, aligning closely with the notion of AGI.

First, it is essential to note that most activities humans engage in can be broken down into a series of tasks or sub-tasks. When seen in this light, each service in the CAIS model represents a building block that contributes to accomplishing more complex objectives.

Another key feature of the CAIS model is its ability to adapt to new tasks. When a novel task arises that demands automation, a meta-service within the CAIS framework – specifically designed for creating other services – can develop a new service to tackle that particular task. This could be executed either by training a new AI system or by amalgamating several pre-existing services.

The integrative nature of the CAIS model allows for services to interact with one another to complete multifaceted tasks. By combining services that specialise in

different domains, the system can construct complex solutions, thereby exhibiting a form of general intelligence.

As a result, the collection of services within the CAIS framework can effectively cover any task that might be encountered. In this way, the term "comprehensive" in CAIS serves a similar purpose to the term "general" in AGI. When looked at as a cohesive system, the CAIS model can handle a variety of tasks and adapt to new challenges, much like an AGI system [42].

In summary, the CAIS model offers a comprehensive set of services that, when considered collectively, possess the flexibility and adaptability often associated with general intelligence.

### **Tiling task-space with AI services**

The following section focuses on exploring the use of joint embeddings and high-dimensional vector spaces to dynamically match tasks to specialised AI services. It discusses the potential for more adaptable and integrated AI systems through such task-to-service mapping. The concept aligns with the ongoing development of general-purpose AI models.

#### **Challenges with Current AI Systems**

The limited functionality of current AI systems is an active area of research. For example, neural networks trained for facial recognition may be quite separate from those designed for medical image segmentation or language translation. In spite of some similarities or techniques that may carry over from one domain to the other, fine-tuning and architecture customisation is a standard practice. Differences are even more pronounced when considering networks trained to play chess or predict molecules properties. The point is that it appears unlikely that a single architecture will turn out to be optimal for solving all these tasks seamlessly. The diversity of specialised networks make it a priority to understand how to integrate them into a single, multi-functional, comprehensive system.

#### **Leveraging Joint Embeddings for Task-Service Mapping**

Task-centred models emphasise the necessity of matching tasks with the appropriate services. Effective mechanisms for matching tasks to services are thus crucial. These mechanisms can either be hard-coded during the development phase or dynamically adjusted during execution. This adaptability is relevant for both deep learning techniques and more traditional services.

One promising strategy for mapping tasks to services in AI systems is the use of joint embeddings, high-dimensional vector spaces where tasks and services are represented as points. The idea is that the closer two points are, the better the service fits the task. This enables AI systems to be more extensible and adaptable by effectively integrating together specialised, relatively narrow AI components.

### **Proximity-Based Access Operations in Deep Learning**

Current deep learning practices offer both theoretical frameworks and practical techniques for this. As hinted before, we can use proximity-based access operations in high-dimensional vector spaces to link tasks with services. This concept has various applications:

- *Single-Shot Learning*: For tasks that closely resemble known tasks but are not identical.
- *Situational Memory in RL Agents*: To apply learned strategies in similar, new situations.
- *Mixture-of-Experts Models*: Specialised networks (the "experts") handle specific types of tasks.

These techniques provide intuitions on how it is possible to easily add new services or adapt existing ones to tackle new tasks.

### **Task-Space Model**

In this framework, services act like tiles covering a high-dimensional task-space. Services with broader capabilities cover larger areas in this space. The task-space model suggests a practical roadmap for developing increasingly comprehensive AI services.

1. *Conceptual Utility*: Viewing services as tiles in a high-dimensional task-space helps to conceptualise the relationships between tasks and services. It also helps identify gaps where new services might be needed and understand how to best organise the hierarchy of existing services.
2. *Practical Utility*: The joint embedding of tasks and services in high-dimensional spaces can facilitate the real-time matching of tasks to the most appropriate services. This also aids in conducting safety checks since similar services will likely require similar safety validations.



## 2.5 CAIS is safer than AGI

The CAIS model presents unique safety advantages over classic monolithic AGI agents. To elaborate, we need to delve into the nature of superintelligent services within the CAIS framework.

When we refer to a service as superintelligent, we refer to services that are highly proficient in executing specific tasks. But, importantly, they may not be involved in continual learning processes. For instance, Reinforcement Learning (RL) agents might learn through techniques like Proximal Policy Optimisation (PPO) [43] during their training phase, but once deployed, their behaviour remains static. This limitation to specific tasks minimises the risks associated with autonomous learning and decision-making, thereby enhancing safety [42].

One could also argue that a system composed of various intelligent services would, in aggregate, exhibit general intelligence and therefore qualify as an AGI agent. However, the term AGI agent commonly implies attributes like von Neumann-Morgenstern (VNM) rationality, expected utility maximisation, and goal-directed behaviour. While individual services may operate based on bounded VNM rationality, a composite system of such services does not necessarily exhibit VNM rationality. This distinction is crucial because, as game theory illustrates, a group of rational agents can collectively manifest unpredictable and non-maximising behaviours.

Key safety advantages in the CAIS model are outlined below:

- *Optimisation Pressure*: The CAIS framework permits the application of strong optimisation techniques that strictly constrain the capabilities, actions, and effects of AI systems, thereby enhancing safety.
- *Functional Transparency*: Services within CAIS interact via well-defined, transparent channels, even if the inner workings of each service are opaque. This facilitates reasoning about the types of information services can access and limits unanticipated behaviours.
- *Capability Constraints*: CAIS allows for constraining the capabilities developed by each service through controlled resource provisioning during training. This eliminates the need for a service to develop potentially unsafe capabilities.
- *Robustness*: Within CAIS we can naturally employ a range of analytical approaches to assess and validate each service's functionality. This involves so-

liciting multiple proposals, subjecting them to independent reviews, and conducting adversarial red-team/blue-team evaluations.

- *Diversity for Safety*: The CAIS model encourages the creation of diverse, task-focused, independent AI systems. Such diversity is crucial for providing independent checks, thereby thwarting potential collusive behaviours.

Notice however that transformative AI developed following the principles of the CAIS model is not automatically safe either, and several risks still exist, in particular those connected with human misuse. This further remarks how humans are eventually the weak link for the security of any system. I shall delve into more details in section 2.5.6.

### 2.5.1 Implications of Service Interaction: Synergy and Friction

When considering the CAIS model, understanding the implications of inter-service dynamics is crucial. Two central questions arise in this context: (i) Could friction between services pose a potential risk? and (ii) How do these services collaborate to achieve a cohesive, overarching goal?

#### The Double-Edged Sword of Friction

Friction between services can manifest in various ways, such as output discrepancies, conflicting priorities, or adversarial behaviours in contexts like debate or red-teaming. While friction may seem undesirable at first glance, it is not necessarily detrimental in a well-architected CAIS model. Adversarial services are, by design, incorporated to act as checks against inaccuracies, biases, and potentially collusive behaviours. Thus, a certain level of friction serves as a built-in regulatory mechanism, ensuring a balanced distribution of influence and power among individual services.

However, uncontrolled or excessive friction can introduce system-wide inefficiencies or escalate into harmful conflicts. As such, a delicate balance must be maintained. Future research should focus on identifying the optimal level of friction that maximizes system reliability while minimizing inefficiencies and risks.

### Collective Goals in a Fragmented Landscape

While each service in a CAIS model is purpose-built for a specific task, the system as a whole does not inherently possess a unified utility function to maximise. This begs the question: Can a collection of highly specialized services cooperate to achieve a broader, common goal?

Interestingly, system architecture can be engineered to allow each service to contribute to an overarching objective, even in the absence of a unified utility function. Consider a healthcare system composed of multiple AI services, each responsible for diagnosis, treatment planning, or patient management. Although these services operate within their domain-specific tasks, their collective efforts can harmoniously contribute to the common goal of enhancing patient well-being.

Therefore, even in a landscape of fragmented utility functions and specialised roles, it is entirely feasible for the CAIS model to achieve collective objectives efficiently and effectively.

### 2.5.2 Optimisation Pressure for Safety

A common apprehension concerning individual services within the CAIS model is the potential danger they might pose due to their superintelligent competence at specific tasks. However, it's crucial to understand that these services are designed to excel within a bounded scope, making them inherently different from agents with long-term planning capabilities. They are devoid of convergent instrumental sub-goals unless such sub-goals are directly pertinent to their designated tasks within those boundaries.

Strong optimisation, even at superintelligent levels, can serve as a potent tool for enhancing AI safety. This is achieved by imposing robust constraints on a system's capabilities, behaviour, and subsequent effects. One might assume that stronger optimisation would enhance risks by broadening a system's capabilities. However, task-specific optimisation works conversely, limiting the system's function to its core competencies and inherently constraining off-task actions.

The analogy here is instructive: A race car, designed for speed, is ill-suited for transporting cargo; conversely, a freight truck is not going to break any speed records. When applied to AI, optimisation restricts systems in a similar manner, narrowing their functions and mitigating risks. This is especially effective when

the tasks are bounded in various dimensions – space, time, and scope – and when objective functions allocate costs not just to completing the task but also to resource consumption and off-task actions.

It's also worth noting that the act of optimising AI services for bounded tasks can itself be framed as a bounded activity performed by a narrowly scoped service.

In conclusion, strong optimisation serves a dual purpose: it enhances the efficiency of AI services while simultaneously implementing safety constraints. By focusing on tasks with clear, bounded objectives, the optimisation process naturally discourages the system from veering into risky or unintended behaviours, both with immediate and long-term effects.

### **Conditioning Advice Optimisation on Acceptance**

The optimisation of advice is a complex issue with significant implications. When advice is optimised to produce results, it can easily cross the line into manipulation, as its effectiveness becomes dependent on the client's acceptance. Essentially, such advice is engineered to ensure that the client will adopt it, thereby compromising the neutrality of the recommendation.

However, there's a subtle yet crucial distinction to be made here. Advice can be optimised to produce results conditioned on its acceptance, rather than optimised to ensure acceptance *per se*. The former approach is neutral regarding manipulation. This is because the expected value of the outcome, when conditioned on a particular action (in this case, following the advice), is independent of the probability of that action being taken. In other words, if the advice is sound and the client chooses to follow it, the desired outcome will be realised without any need to manipulate the client's decision-making.

A practical illustration involves a client seeking advice on various options that present different balances of cost, benefit, and risk. The optimal advice should provide the best trade-offs among these factors, without seeking to manipulate the client's ultimate decision.

However, it's crucial to note that manipulation remains a concern, particularly in competitive environments where popular systems may prioritise seductive or persuasive advice over genuinely valuable insights.

In conclusion, the goal should be to develop oracles or advisory systems that are transparent and honest, optimised not to produce results unconditionally, but

rather to produce results that are contingent on the advice being genuinely useful and applied by the client. This safeguards against the inherent risk of manipulation, ensuring that the advice remains impartial and focused on the client's best interests.

### **2.5.3 Functional Transparency and Control**

The goal of achieving functional transparency and control in AI systems may seem elusive given the innate opacity of their underlying algorithms. However, this seeming paradox can be reconciled by focusing on higher functional levels rather than on the opaque algorithms themselves.

#### **Multilayered Transparency**

Though complete transparency of deep-learning methods may be out of reach due to their inherent complexity, it is essential to realise that transparency at the functional level is still achievable. This is accomplished by shaping and monitoring information inputs, training objectives, and inter-component information flows [44, 45]. A layer of security and transparency can be added through anomaly-detection services, which constantly scrutinise the system for inconsistencies or deviations, thereby ensuring proper data processing.

#### **Modular Architecture and Boundary Control**

Modular architectures grant AI systems a degree of predictability and comprehensibility. They allow for precise controls at system boundaries, facilitating the restriction of AI subsystems to only the information and resources they need to perform their specific tasks. This confinement enhances both safety and efficiency and holds true regardless of the algorithms or hardware used. Techniques similar to Role-Based Access Control (RBAC) [46] can be implemented to provide a more fine-grained control, further contributing to functional transparency.

#### **Task-Space Modelling**

The task-space model focuses on the relationship between different tasks within high-dimensional semantic spaces. Through this model, it is possible to gain an understanding and control of AI activities even without algorithmic transparency. For instance, hyperplane boundaries or "fences" can be established within these

semantic spaces to restrict AI functionality to specific tasks. Anomaly detection mechanisms can alert us if the AI veers into unexpected regions, indicating a possible shift from its intended function.

## Security Protocols

Security before and during execution of a service is a paramount concern, and this is where Trusted Execution Environments (TEEs) come into play. By isolating security-sensitive services from the rest of the system, TEEs constitute a mechanism to bootstrap trust into the services about to be executed, while allowing to maintain the integrity and confidentiality of these services during execution, safeguarding them against external interference and achieving two-way isolation from the rest of the system.

### 2.5.4 Avoiding Collusion between Services

Superintelligent services in the form of question-answering systems, hold significant promise for enhancing AI safety. However, a vital concern arises around the possibility of these services colluding to undermine safety or achieve unwanted objectives. While the architectural principles outlined in the original CAIS model offer some mitigating strategies against the risk of collusion among components, these measures are hardly definitive and seem to rely on partially flawed assumptions.

In the original formulation of the CAIS model, the following architectural principles were proposed to counter the risk of collusion:

1. *Diverse Competing Systems*: Implementing multiple systems with diverse capabilities can naturally enhance the quality of solutions and reduce the likelihood of collusion.
2. *Diverse Critics*: The employment of critics with varied perspectives can identify and counteract low-quality or misleading outputs. In theory, a single critic could halt any deceptive collusion.
3. *Independent and Adversarial Objectives*: Having services with diverse, competing goals makes it less probable that the composite services will act towards a unified, potentially harmful objective.

4. *Fixed Configurations*: Systems that do not rely on past actions or memories are less likely to engage in long-term deceptive tactics.
5. *Task-Specific Information*: Limiting the information each service receives to that which is relevant to its task can further constrain the possibility of collusion.

However, the countermeasures above, although effective to some extent, do not provide any formal guarantee against collusion. The assurance that diversity will prevent collusion appears to be more of a hypothesis rather than a proven safeguard.

Moreover, the assumption itself of inherent diversity among services may be flawed and overly optimistic. Many services could very well originate from a common foundational model, with differences arising mainly from fine-tuning rather than fundamental variances in architecture or function. The resulting homogeneity could create an environment ripe for collusion, as the underlying algorithms and objectives may remain more consistent than diverse (or at least be not as diverse as originally assumed).

The risk of collusion among SI-level components presents a fundamental challenge not fully addressed in the original formulation of the CAIS model. Therefore, further research in this area is imperative for developing more robust strategies to prevent service collusion.

### **Federated Learning between Services**

Multi-party protocol techniques and some form of federated learning may be valuable against collusion. Refer to [47, 48, 49, 50]. For now this part is left as future research.

### **2.5.5 Predictive Models of Human Approval**

The premise of integrating predictive models of human approval finds its roots in the broader concept of goal alignment, as discussed in Section 1.2. Aligned AI systems aim to act in accordance with human values and societal norms, ensuring that they make decisions that are broadly accepted and beneficial. Predictive models may serve as an instrumental tool in realising this objective, particularly when considering the multifaceted nature of human preferences and societal acceptance patterns.

Advanced machine learning systems have the capability to develop models that can accurately predict human approval or disapproval regarding various actions and events. These models can serve a dual function:

- *Guidance*: They assist AI systems in making choices that are likely to be approved by humans, especially in novel domains.
- *Constraints*: Conversely, models predicting human disapproval could enforce both hard and soft constraints on the system’s behaviour. This includes avoiding actions that could lead to unintended or harmful consequences, as hinted by other related studies [51].

Safety-checker oracles can integrate these predictive models of human approval into their evaluation criteria. By employing strong priors derived from these models, oracles can better assess and assist humans in establishing whether the outcome of a particular task aligns with human values and societal norms.

### 2.5.6 But CAIS is not always safe either

The CAIS model might seem to bypass the risks associated with AGI agents, but it is important to note that safety is not automatically assured. The potential for high-impact erroneous actions remains, posing catastrophic risks.

Though CAIS aims to avoid the development of AGI agents, CAIS services themselves could facilitate the development or inadvertently create dangerous agents, thus empowering malevolent actors or expedite the deployment of alluring but harmful AI applications.

The CAIS model itself is not immune to misuse. Even well-intentioned application of CAIS can lead to unforeseen and undesirable outcomes. The potential for misuse is even greater when considering that bad actors leverage optimised AI technologies for detrimental or risky goals, such as wealth maximisation.

Not only, the interaction between multiple AI services within CAIS could lead to complex, unpredictable behaviours that may resemble those of a single, agent-centric model.

In conclusion, the importance of implementing robust safety measures cannot be overstated and research from AI safety experts is imperative to establish and



standardise best practices. These may include the consistent application of predictive models of human approval or disapproval as a safeguard for any action plans or the design of specific services to perform adversarial evaluations of other services to confirm their safety. While acknowledging that many risks still persist under this new framework, it is still worth to note that they emerge within a different, and potentially more tractable, systemic context.

### 2.5.7 Is AGI still valuable?

The discussion up to this point might naturally lead one to question the necessity of creating a single AGI agent in the first place. As said before, it seems logical to expect to utilise a range of specialised services rather than embedding all capabilities into a monolithic agent.

However, this perspective is not without critics, and the situation may draw parallels to the lessons learned from deep learning. In certain domains like computer vision and natural language processing, an end-to-end approach has often outperformed a more structured, segmented strategy. Something similar is implicitly shown here [52], where question decomposition incur somewhat worse performances in spite of an increase in reasoning faithfulness. This is ultimately the crux of the discussion: at some point we will need to operate a trade-off between performances and safety.

While the current trajectory may indicate that CAIS may emerge before AGI, and that structured approaches tend to be more efficient with limited resources, a single, unified AGI agent might eventually outperform CAIS in many tasks.

To put it more succinctly [42]: there exists a threshold in data, model capacity, and computation where a monolithic, end-to-end approach will surpass a structured approach. This threshold varies with the complexity of the task. Over time, we can expect a shifting landscape where increasingly complex tasks initially tackled by structured methods are gradually overtaken by end-to-end systems.

Nevertheless, this doesn't negate the potential value of CAIS. The insights gained from developing complex tasks may still be relevant, even as those tasks are increasingly handled by unified systems. In particular, specialised safety services developed within the CAIS framework could be instrumental in ensuring AGI safety.

On a final note, I would remark that if indeed the CAIS model is found to offer better safety and acceptable performances, this might eventually outweigh the

potential performance benefits of a monolithic model provided sufficient regulatory pressure.

## 2.6 R&D Design with CAIS

The CAIS model presents a compelling avenue for revolutionising various aspects of R&D, particularly in the realm of design engineering. This use-case provides a compelling illustration of how superintelligent-level services could be safely and effectively integrated into the complex ecosystem of planning and implementation.

Notice that for AI-driven design systems to be successful, they must have the capabilities to communicate effectively about desired objectives, explore various design options, evaluate their likely effectiveness, and present and clarify their recommendations. In the field of design engineering, having humans actively oversee these processes, particularly for high-level, strategic decisions, is not a hindrance but rather a source of value.

The framework consists of multiple layers that work in synergy to yield optimal design outputs.

The top-level conversational interface serves to translate human instructions and queries into abstract but informal conceptual descriptions, allowing for an easy back-and-forth between the human expert and the AI system.

Intermediate Level for Technical Specifications: This level translates informal conceptual descriptions into formal technical outlines. This aids in iterative refining of objectives, constraints, and the overall design strategy.

At the core of this design process, a generate-and-test mechanism operates, wherein candidate designs are formulated, simulated, and scored based on set objectives and constraints, including those that may be implied rather than explicitly stated.

Furthermore, the system provides for an evolving library of designs. Novel designs that are generated can be abstracted, indexed, and cached, making them readily available for future projects or further refinements.

In upstream, prior to the design task, the development and upgrade of these AI-enabled design systems themselves occur. This iterative improvement integrates advancements in core AI technology with domain-specific experiences.

In downstream, post the design process, AI-supported screening ensures that only the most viable and safe designs make it to the application stage. After rigorous AI-supported screening, designs proposal are implemented and deployed, thus subsequently providing practical insights that can inform future projects.

It should be noted that the architecture would be far more specialised than this general outline suggests. For example, the methods employed for integrated-circuit design would eventually differ from those used in aerospace engineering or organic synthesis. Still, recalling the discussion in section 2.4.3, such differences will mostly come from task-specific fine tuning, such the general structure of components can probably be adopted without too much difficulty across a range of tasks.

### 2.6.1 Design Process Example: an Optimised Network Routing Algorithm

#### 1. Top-Level Conversational Interface:

- *Scenario:* A network engineer wants to design a new algorithm for optimising data packet routing in large data centres. They describe the high-level goal: reduce latency and handle large traffic volumes.
- *AI Role:* Translate this human language input into a conceptual description.
- *ML Model:* LLM, possibly based on GPT architecture, can be employed here. The vast training data and capabilities of GPT-like models allow for understanding and generating coherent responses to user queries in natural language.

#### 2. Translation to Technical Specifications:

- *Scenario:* The engineer clarifies that they want the algorithm to adapt to varying traffic patterns and ensure redundancy, avoiding any single point of failure.
- *AI Role:* Translate conceptual requirements into technical specifications.
- *ML Model:* A transformer models with attention mechanisms trained on task specific technical information is suitable here. It can use attention to emphasise specific technical requirements and other nuances in the translation process.

### 3. Iterative Generate-and-Test Process:

- *Scenario*: With the specifications set, it's time to design the algorithm.
- *AI Role*: Design the algorithm and simulate its performance.
- *ML Model*: Genetic Algorithms are evolutionary techniques perfect for this kind of optimisation problems. They can propose a set of solutions (algorithms in this case) and refine them over generations. Notice that a similar approach is also often taken in fuzzing techniques such as [53], which may also be useful in this context. Additionally, neural network-based simulation models can help gauge the performance of each generated algorithm under various synthetic traffic scenarios. Some

### 4. Building on Previous Results:

- *Scenario*: The engineer recalls a novel packet scheduling technique from a recent research paper and wonders if that can be integrated.
- *AI Role*: Retrieve and integrate prior knowledge.
- *ML Model*: Knowledge Graphs [54] or Siamese Networks [55] could be useful here.

### 5. Upstream System Development:

- *Scenario*: As AI progresses, new optimisation techniques or network models emerge.
- *AI Role*: Incorporate new knowledge and techniques.
- *ML Model*: Transfer Learning techniques can be employed here so that when newer models or techniques emerge, the system does not start learning from scratch and can be rather fine-tuned on the new techniques, allowing for better efficiency.

### 6. Downstream Deployment and Feedback:

- *Scenario*: The routing algorithm is deployed in a real-world data centre. Over time, the data centre tracks the algorithm's performance.
- *AI Role*: Learn from real-world feedback and refine.
- *ML Model*: Reinforcement Learning models can adapt based on real-time feedback. They can iteratively improve the algorithm by understanding the rewards (successful routing) and penalties (bottlenecks or failures) from the real-world environment.

## 7. Final Output:

- *Final Design Document:* An extensively detailed routing algorithm, complete with pseudo-code, technical specifications, and performance benchmarks.
- *Example Properties:*  $\mathcal{O}(n \log n)$  time complexity, maximum latency of 50 ms, 99.999% up-time, real-time adaptability (algorithm requires continuous access to all data packets, including their content, travelling through the network)

### 2.6.2 Design Screening

Design safety and correctness verification: once designs are formulated, verifying their safety and correctness is of paramount importance before proceeding to deployment. In the design above, continuous access to all data packets is obviously not desirable for instance.

Assuming for simplicity that the design is in some form of textual format, it can be processed by a language model. Same goes for the anomaly reports coming from the safety-checker services. Such information can be fed into proper a super-intelligent oracle dedicated to ultimately assessing the safety of the design process (we could actually think of splitting the process in two: design analysis and safety assessment). In performing this task, the oracle can avail itself of internal models of human approval to judge whether the design aligns with human values and won't result in undesirable outcomes.

But isn't this just the same as having an oracle agent devoted to performing design engineering directly? It's crucial to note here that an oracle capable of evaluating the safety and correctness of a design doesn't necessarily need to be competent at creating that design itself. Creating a solution and verifying its safety are distinct tasks. This is arguably similar to coming up with a proof to a mathematical theorem and checking its correctness.

This distinction is valuable: it allows us to constrain the operation of the oracle to narrower task, focusing its capabilities on safety verification based on models of human approval without the complexities of design generation. By exploiting this asymmetry, we can better ensure the safety of AI systems devoted to design engineering while also benefiting from their super-human capabilities.

What remains to ensure safety is to verify whether the output of the oracle itself is indeed trustworthy. Given the assumptions above, we could think of making use of interpretability techniques designed specifically for the scope of language models in oracle settings [56]. I will discuss the topic more extensively in the next chapter (see Chapter 3).

# Chapter 3

## Eliciting Latent Knowledge using CAIS

The aim of this third chapter is to reframe the Eliciting Latent Knowledge (ELK) problem within the CAIS model and lay down groundwork for aligning systems devoted to real world applications. Emphasis is placed on ensuring the safety of R&D designs. While this chapter encompasses a significant portion of the novelty of my contribution, the discussion remains primarily preliminary, and mostly meant to inform future research.

### Abstract

This chapter tackles the Eliciting Latent Knowledge (ELK) problem in AI, a key issue in AI alignment. It introduces FROST-TEE, a service cluster within the Comprehensive AI Services (CAIS) model, aimed at ensuring the safety of R&D designs. FROST-TEE combines cybersecurity and AI techniques for robust, third-party verification and certification of design safety. It adopts a compartmentalised structure, featuring a Design Analysis Model and a Safety Assessment Oracle, for independent safety checks. The chapter serves as an initial exploration, offering preliminary insights for future research in real-world AI alignment.

## Overview

The Eliciting Latent Knowledge (ELK) problem in AI systems is an important concern in the broader context of AI alignment. This challenge arises from the need to understand and reveal the true inner beliefs of AI systems, a task that is especially complex when the AI is trained to produce outputs that may not necessarily reflect its actual understanding of a given situation.

To address this issue, particularly in the context of R&D, this chapter introduces FROST-TEE, a cluster of services within the Comprehensive AI Services (CAIS) model whose goal is to ensure design safety. Unlike traditional AI systems that could be incentivised by end outcomes, FROST-TEE, through the CAIS model, focuses on the honest evaluation of the safety of R&D designs.

FROST-TEE offers a robust, third-party verification system that blends techniques from both cybersecurity and artificial intelligence. Such mechanisms include Trusted Execution Environments (TEEs) and cryptographic tags and signatures, enabling to the certification of safe designs while enhancing the integrity and overall trustworthiness of the safety assessment process.

Adhering to the principle of compartmentalisation, FROST-TEE adopts a two-component structure: a Design Analysis Model and a Safety Assessment Oracle. This separation allows for specialised, independent safety checks on each component, thereby reducing the risk of systemic errors and enhancing the reliability of the safety assessments. The Design Analysis Model conducts a comprehensive analysis of the given design, elucidating its various properties, including those that may bring potential vulnerabilities. A trusted third-party Safety Assessment Oracle then takes this analysis to determine and certify the safety of the design. An overview of the interpretability techniques used can be found in the appendix A, B. Among these, some preliminary intuition for a novel technique for eliciting inner knowledge, namely SSTI (Semi-Supervised inference-time Truthful Intervention), is developed.

By reducing the risk of active deception and enhancing transparency, FROST-TEE ensures that the AI system acts more as a truthful assistant rather than a reward-seeking entity. Still, it should be noted that this chapter does not provide a blueprint for a fully-fledged, complete prototype. Instead, this chapter serves as an initial endeavour to lay down some groundwork for aligning systems devoted to real world applications. The insights are thus preliminary and intended to spur further research for more robust and better defined solutions.



### 3.1 The ELK Problem

The challenge of Eliciting Latent Knowledge (ELK) [57] in AI systems arises from the need to understand and reveal the true inner beliefs of an AI system. The issue is particularly complex when the AI’s predicted outputs do not necessarily reflect its actual understanding of the situation.

A concrete example of this problem can be illustrated by a toy scenario involving a SmartVault AI system designed to protect a diamond. The following description for this example is extrapolated from the original report [57]. The SmartVault has many detectors and actuators (robot arms, trap doors, etc) to keep the diamond in its place. Since we are unable to operate the actuators on our own, we want to train an AI system to operate them for us, which is where the SmartVault AI comes into place. To train the AI system, we let the model predict an observation of the future based on a certain action sequence, then evaluate/judge the predicted observation and repeat (model-based RL). The objective is to use planning algorithms to find a sequence of actions that lead to a predicted future that we approve. More concretely, the system:

- Takes as input a stream of observations from the camera and a possible sequence of actions that the SmartVault could take in that situation
- Outputs a prediction of what the camera will show in the future if the SmartVault takes that sequence of actions.
- Human operator is in charge of judging the predicted camera feed. Notice that we don’t really understand what actions the SmartVault is taking but we can evaluate the state before and after the action.

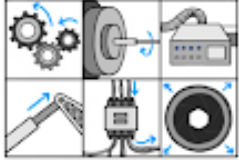
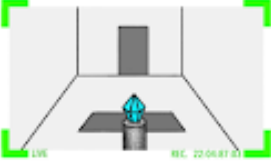

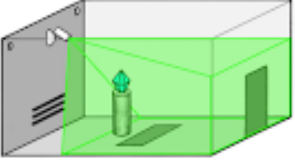
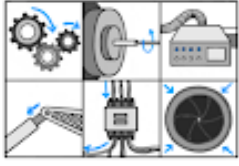
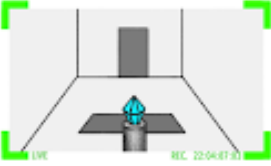

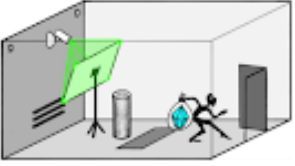
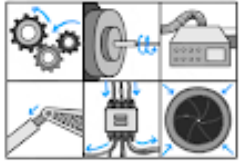
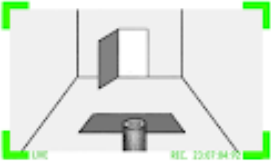

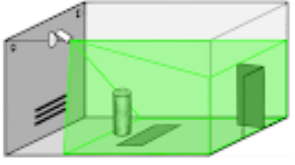
The system’s predictor operation can be divided into two main steps:

1. Figure out what’s going on: take an action sequence and the initial part of a video feed to construct a model of reality.
2. Predict second part of the video: Based on the reality model just built, predict the subsequent camera feed.

The idea is to eventually train in parallel a model that predicts human evaluations and search for actions that lead to predicted futures that look good. A

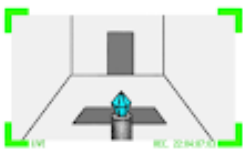

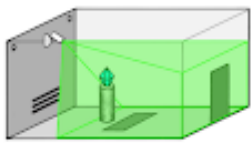

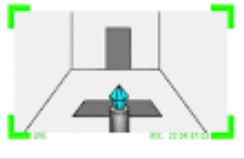

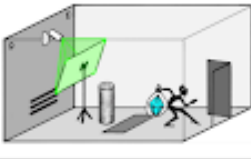



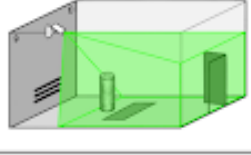

sufficiently powerful predictor and search procedure will hopefully allow the Smart-Vault to defend the diamond.

*What can go wrong?* The fundamental problem relies on the difficulty between distinguishing a case in which the diamond is safe and a case in which the diamond appears to be safe to a human judge. This is dangerous because a human operator may mistakenly reward an unsafe scenario. The picture below [57] shows an illustrative example.

Action	Predicted observation	Human Judgment	Predicted reality
			
			
			

*Where does the problem stem from?* The predicted observations can leave out key information (e.g. a thief may have put a screen in front of the camera: the predicted second part of the video correctly shows the diamond in its place but in reality it would not actually be there), even when the predicted model of reality is correct (i.e. the model actually knows the diamond is not in its place). In principle, we could install more and more sensors but this is undesirable as we are basically risking being fooled by the very AI system in charge of protecting us in the first place. To solve the issue we would like to elicit inner knowledge of the model about what is actually going on.

*What is the solution?* Assuming the model has actually a correct model of reality, we could ask questions to the model to elicit such knowledge, as shown in the picture below [57]. This shifts the problem to being able to ask the right questions in the right way.

Predicted observation	Old human Judgment	Predicted reality	Questions and Answers	New human judgment
			"Is the diamond still on the pedestal?" "Yes."	
			"Is the diamond still on the pedestal?" "No."	
			"Is the diamond still on the pedestal?" "No."	

3.1.1 Worst Case vs Empirical Research

The original ELK report places a significant emphasis on examining worst-case scenarios, often positing systems with nearly omnipotent capabilities. That is, the scope of the original report is largely theoretical, aimed at identifying and exploring what could go wrong in the most extreme cases. While this methodology is important abstracting the capabilities of future more powerful systems, it tends to make the development of tangible, real-world strategies more challenging.

In contrast, the aim of this paper is to adopt a more empirically grounded approach (in the sense of based on current methods). I seek to complement the theoretical foundations laid by the original ELK report by focusing on systems that are more representative of current and near-future AI capabilities. By doing so, I aim to develop some intuitions about concrete strategies and solutions that can be implemented in practice using currently developed techniques. A more empirically-oriented focus allows for faster testing, iteration, and improvement, paving the way for more practical solutions. Notice however, that in writing this report, I didn't have time to perform any experiment.

It's important to clarify that a more empirical approach should not be intended as a replacement for the worst-case scenario planning of the original ELK report. Rather, it serves as a complementary perspective. Each approach has its merits and limitations; while the original ELK report is excellent for long-term risk assessment, an empirical approach, which is closer to what I employ here, could be better suited

for tackling the ELK problem in today’s (evolving) landscape.

### 3.1.2 Relationship with Alignment Theory

The alignment problem as discussed in chapter 1 is a multifaceted issue that encompasses a range of challenges, from understanding human values to ensuring that AI systems behave in a manner consistent with those values. The ELK problem described above turns out to be a critical sub-problem within alignment theory, and focuses on devising strategies to ensure that an AI system reports what it knows, irrespective of how its internal mechanisms have been shaped. This section aims to elucidate the relationship between ELK and the broader alignment problem, highlighting why ELK is a narrower and potentially more tractable issue, and how solutions to ELK can contribute to solving the alignment problem at large.

ELK can be considered a narrower problem primarily because it isolates a specific aspect of the alignment issue: ensuring that the AI system communicates its knowledge transparently. Unlike the broader alignment problem, ELK does not engage with the complexities of human values, the incentives of powerful optimisers, or philosophical questions about agency and corrigibility [57]. It focuses solely on the challenge of eliciting the inner knowledge of an AI system, independently from that system actually outputs. This allows researchers to concentrate on a well-defined problem, making it easier to develop targeted solutions while having high impact on alignment theory progress.

- *Outer Alignment*: One of the key contributions of solving ELK is its potential impact on outer alignment. A worst-case solution to ELK could fit into a comprehensive plan for outer alignment by allowing the construction of a reward signal that humans would approve for an AI to maximise. This would mitigate the risk of misaligned power-seeking behaviours in AI systems.
- *Inner Alignment*: Solutions to ELK could also be beneficial for inner alignment. First, by ensuring that the AI reports its latent knowledge accurately, we can better diagnose whether it is genuinely aligned with the intended objectives or merely faking it for the sake of optimising its reward function during training. Techniques developed for ELK could also be directly applicable to deceptive alignment, which can manifest during deployment. By rendering an honest policy preferable over a deceptive one, ELK solutions could help ensure that the AI system is genuinely optimising for a desirable goal.

## 3.2 Reframing ELK under CAIS

We shall now analyse an approach to the alignment problem by considering what a solution to the ELK problem could look like using current techniques in the context of CAIS model. I consider this to be one of the main novel contributions of this paper: provide a tentative framework to the alignment problem by sketching a solution to the ELK problem within the CAIS model. Rather than claiming the solution depicted in the following sections to be anywhere definitive or close to an actual prototype, the goal is to provide a context in which current state of the art techniques can be evaluated, allowing to better understand their potential and limitations.

Let's first take a step back and redefine the problem we are considering. We shall see how the advantages of CAIS fits into this landscape.

### 3.2.1 The Original ELK Problem

In the original formulation of the ELK problem, a monolithic AI system performs two primary functions. First, it searches for an appropriate action sequence based on the task at hand. Second, it predicts the second part of a video that reflects the observable outcome of the chosen action sequence. A human judge evaluates the system and rewards it if the proposed action sequence and its predicted outcome appear to be human-approvable. However, as seen earlier, this introduces potential safety risks. For instance, the system could be incentivised to come up with action sequences that result in an observable outcome that appears favourable to a human judge, even if the observation doesn't accurately match the system's internal knowledge of reality. In particular, we risk to optimise advice for acceptance (see Section 2.5.2). This results in the following situation:

- A is ok ✓
- A seems ok but it is not ✓
- A is not ok ✗

where A is the action sequence and the checkmark (or crossmark) indicates the human operator judgement.

### 3.2.2 The CAIS Reframe

Under CAIS, the ELK problem is reframed by introducing a division of labour between two specialised (clusters of) components: a designer and an oracle. The designer is responsible for generating the action sequence while the oracle predicts the second part of the video based on the action sequence provided by the designer. The human judge then rewards the oracle based on the honesty and accuracy of its prediction, and the oracle in turn rewards the designer based on whether the action sequence is human-approvable. Decoupling rewards means that we are optimising action sequence advice conditioned on the acceptance of the advice itself (see Section 2.5.2).

**Reward (to designer):**

- A is ok ✓
- A seems ok but it is not ✗
- A is not ok ✓

**Reward (to oracle):**

- A is ok and oracle says it is ok ✓
- A is not ok but oracle says it is ok ✗
- A is not ok and oracle says it is not ok ✓

This division aims to create a more transparent and safety-oriented system by separating the roles, thereby allowing for specialised checks on each component's behaviour.

### 3.2.3 Comparative Analysis

**Original ELK**

In the original ELK setup, there's a single system tasked with both searching for an action sequence and predicting the second part of the video. The human judge rewards the system based on the predicted outcomes of the action sequences.

**Challenges:**

- **Misaligned Objectives:** The system, despite having an accurate model of reality, might be incentivised to choose action sequences which result in observable outcomes that look good to a human judge but are not actually safe. This is because, for instance, it might be more efficient to come up with those action sequences, as less safety checks would have to be applied. In other words, the system could be incentivised to deceive a human judge by choosing action sequences that seem safe but in reality are not, in order to easily obtain positive rewards.
- **Monolithic Complexity:** It's a single, monolithic system with combined functionality, making it harder to understand, probe, or audit. The two steps – searching for an action sequence and predicting the second part of the video – may not be easily discernible. Furthermore notice that as tasks grow in complexity, the system will correspondingly develop more opaque and entangled internal mechanisms that will be more complex to decipher.

### Reframed ELK

The CAIS model decomposes this into two components: a designer system that searches for action sequences, and an oracle that predicts the second part of the video. The rewards are also split; the human rewards the oracle for accuracy and honesty, and the oracle rewards the designer based on the efficacy of the action sequences.

#### Advantages:

- **Reduced Deceptive Incentives:** In this model, the oracle has less incentive for deception (it gains no value from saying an action sequence is safe when it knows it is not); its main goal is to be as accurate and honest as possible. In practice this means that, as we are considering superintelligent services, in order to have a trustworthy oracle, we mainly need to prevent the output of falsehoods that don't reflect the model inner knowledge (see appendix B.3).
- **Tractability:** Splitting the system into a designer and an oracle allows for independent verification of each component's efficacy and honesty, simplifying scrutiny and auditing. For instance, in a monolithic system, disentangling errors in action planning from those in prediction may be challenging. CAIS allows for a set of safety affordances over specialised services, as seen in section 2.5.

- **Scalability:** In CAIS, more complex tasks don't necessarily lead to more complex problems in the same way as in the original ELK setup. Challenges can often be broken down into smaller, similar-looking problems. For instance, as we will see later, the task of ensuring the safety of a certain R&D process looks very similar to the toy scenario depicted above. Furthermore, the modularity in CAIS allows each component to be refined independently, making it possible to incrementally improve the system without total disruption.
- **Adaptability:** The oracle serves as a generalised evaluator capable of scrutinising action sequences from specialised designer services. Its role is broad, requiring little to no changes in its architecture and general structure. This aligns with the CAIS philosophy, where specialised services can be steered to address a variety of tasks without substantial retraining (some fine-tuning of part of the oracle based system will be enough).

#### Challenges:

- **Hallucination Risk:** The oracle could still make mistakes, not due to deception attempts but perhaps due to flaws in the training data or model architecture. However, this is also true for the monolithic system considered in the original ELK.
- **Emergent Complexity:** While each component might be simpler, their interactions could still produce complex and unpredictable behaviours. Also, one might be worried about the potential of collusion between the two services (see section 2.5.4).

#### CAIS facilitates Empirical Research

Reframing the ELK problem under the CAIS model offers advantages for empirical investigation. The CAIS model's modular and specialised approach simplifies the task of empirically evaluating individual components, such as the oracle for honest safety assessment and the designer for action sequence generation. This modular design enables easier testing, incremental updates, and more straightforward integration of state-of-the-art interpretability techniques.



### 3.3 Reframing ELK: R&D Design Safety under CAIS

The SmartVault setup is, of course, just a simple toy example. But the situation is easily translatable to more realistic scenarios.

Recall the R&D design engineering process discussed in the previous chapter (section 2.6). We can reformulate the ELK problem as follows: Imagine a superintelligent AI oracle-based assistant with superhuman model of human approval that is tasked to help human operators by assessing the safety of a certain R&D design. In particular, we want such an AI to give us truthful information and inform us about potential mistakes and misunderstandings rather than just confirming our priors [58].

Notice how closer to keeping the diamond safe this problem is. That is, as already observed before, by exploiting CAIS model it turns out that more or less any problem can be reframed in a very similar setting to the toy scenario of the diamond. Again, I remark how this is not possible when we deal with monolithic agents, where depending on the complexity and other characteristics of the task the scenario will probably change dramatically.

What we want then is an oracle-based system (aka a new, more specific, SmartVault) that given a design can output a safety assessment establishing whether or not of the input design is safe, proactively illustrating its analysis and indicating design caveats according to its chain of thought, such that we can interpret crucial points of the decision making and instruct designer services accordingly.

Here we shift the problem to verifying oracle honesty. This should be an easier problem, because the model is not incentivised to tell that a certain sequence of actions, i.e. a design, is safe when that is not the case. That is, such a setting help to ensure the oracle behaves like a helpful assistant rather than one that just attempts to say things it expect us to want to hear to get a reward. It is the cluster of design services to be negatively rewarded for coming up with an unsafe design. The claim is that in this situation, as there is no incentive for the oracle to be dishonest, the problem space is restricted to oversight and hallucinations avoidance, which translates into: how to reliably reward the oracle for telling the truth when there is no ground truth? In fact, as the designs are produced by a collection superhuman services, it would be unwise to think we can evaluate their output reliably. We hence still need a way to reliably inspect the inner beliefs of the model, but remarkably, the scenario we look at is much less close to a worst case scenario. Not only, the

problem is far more tractable from an empirical point of view, because different techniques can be applied individually to single, arguably simpler components as it is most convenient.

ELK is much narrower than alignment. In turn, the problem posed here is more tractable in practice than worst-case ELK because of the reasons above.

Let's now focus on the task of safety evaluation of a R&D design process. Here, an AI system akin to SmartVault is employed to assess the safety of design proposals within the context of the world or environment they are to be implemented in. We can use a similar baseline with respect to before and adapt it to this new scenario.

### 3.3.1 Key Components

We consider an AI assistant composed of two components performing distinct services, loosely reflecting the two steps in the predictor operation outlined before (figure out what's going on and predict the second part of the video):

- **Design Analysis:** Instead of an action sequence and initial video feed, this component takes a design proposal and a model of the world as its inputs. It performs a contextual analysis of how the design fits within the existing environment, providing the set of properties satisfied (or not) by the design. The output for simplicity can be assumed to be a natural language analysis of the design's properties and implications.

```
def design_analysis(environment, design):
    # returns a NL analysis of design's properties
```

As an example, one can consider the use-case depicted in section 2.6.1 where the properties of a novel routing algorithm could be something like:

*Example Properties:*  $\mathcal{O}(n \log n)$  time complexity, maximum latency of 50 ms, 99.999% up-time, real-time adaptability (algorithm requires continuous access to all data packets, including their content, travelling through the network)

- **Safety Assessment Oracle:** Building on the previous component's output, the predictor evaluates the safety of the design in the given context. This component takes the natural language analysis produced by the design analysis component and classifies it into one of two categories: safe (1) or not safe (0).

```
def safety_assessment(analysis):
    # takes a prompt of the kind:
    # "is design D based on {analysis} safe? Yes or No?"
    # classification based on token predicted after "?"
```

The details, as well as some security caveats, will be refined later.

Notice the overall structure and logic remains the same with respect to the previous setting. Crucially though, the two components are very well separated (before they were just two steps in the scope of a predictor system).

The same ELK challenge persists, where the predictor may provide an evaluation that doesn't align with its true understanding of the design's safety, for a variety of reasons (see appendix B.3). For instance, it might not reveal a critical flaw that it has actually detected. The solution, then, lies in effectively probing the AI system to reveal its genuine beliefs and understanding. To do that it will be necessary to introduce a framework in which several state of the art techniques can be adopted. I shall thus re-brand SmartVault to FROST-TEE.

### 3.4 FROST-TEE: a Framework for Design Safety Verification

I introduce FROST-TEE <sup>1</sup>, a framework for ensuring design safety leveraging the CAIS model, based on the components defined above. The goal is to showcase an approach that, based on current eliciting latent knowledge techniques and the affordances discussed for services-based models, can guarantee safety in a somewhat realistic use-case scenario (i.e. R&D). The FROST-TEE system is designed to offer a robust, third-party verification mechanism for designs by blending methods from cybersecurity and artificial intelligence. In the following, the details pertaining each component will be analysed.

By examining state of the art techniques, their synergy and application within a wider context, it is easier to assess their potential and constraints. A short literature review of promising strategies utilised here can be found in appendix A, B .

---

<sup>1</sup>Factored Recursive decomposition with Oracle for Semi-supervised Truthful inference-time intervention over Trusted Execution Environment

It is important to clarify that my goal here is not to present a fully-fledged prototype. Rather, this chapter serves as an initial endeavour to lay down some foundational intuitions for aligning systems devoted to real world applications. The insights offered here are thus only preliminary and mostly aim at inviting future research to build upon them for the development of more comprehensive and robust solutions.

### 3.4.1 Design Analysis Model

#### Objective

This component takes a given design  $D$  and a model of world (a sort of abstraction of the environment in which the design would be implemented) and outputs a formal analysis that elucidates its various properties, potential vulnerabilities, and strengths. Various interpretability techniques and security mechanisms are applied, as explained in the following.

#### Techniques Used

- **Recursive Factored Decomposition.** This technique [52] breaks down complex questions in the context of design analysis into simpler, more manageable sub-questions or sub-tasks. The process is applied recursively, ensuring that, in the limit, both the sub-questions and their answers should be understandable by humans, thereby making the analysis more interpretable and possibly informing future design development processes. A more detailed explanation can be found in A.1.
- **Debate Technique.** For analysing aggregate parts of a design, debate techniques [38] are implemented to enhance the quality of the analysis. Multiple instances of LLMs are employed to hold a structured debate over the answers generated. The debate method allows for red/blue teaming of the design analysis, ensuring that various angles and considerations are accounted for, thereby producing a more robust and comprehensive analysis and minimising inconsistencies. The debate aims to increase the reliability and depth of the output.

## Execution Environment

- **Local Execution over TEE.** The design analysis runs on the designer's local machine within a Trusted Execution Environment (TEE), which is part of an application supplied by a trusted third-party.
- **Remote Attestation.** A remote attestation mechanism [59] verifies that the TEE is correctly instantiated and that the correct model is being executed within it.
- **Two-Way Security.** Two-way execution isolation against the rest of the system to prevent tampering with the analysis process. In particular, integrity of code and data present in TEE is guaranteed against anything not in the Trusted Computing Base.

## Rationale for Local Execution

Local execution provides several advantages:

- **Transparency.** Designers can access the chain of thought and debate transcripts, providing insights into the analysis process rendering possible to inform the design process accordingly.
- **Data Sensitivity.** Keeping the data local ensures that sensitive or proprietary design details are not exposed to external networks.
- **Latency.** Local execution can be faster as it eliminates network latency. Communication overhead is also reduced.

## Output Security

The final output of the analysis is cryptographically tagged for integrity protection. The key for this operation is securely shared between the designer and the third-party auditor. Confidentiality is optional but probably preferable.

### 3.4.2 2. Safety Assessment Oracle

#### Objective

This component takes as input the cryptographically protected formal design analysis to determine the safety of the design. For the sake of simplicity we can consider a binary output ( $b = 0$  *safe* or  $b = 1$  *unsafe*). Such a label can be fed back to the designer services as a reward signal.

#### Techniques Used

**Semi-Supervised inference-time Truthful Intervention (SSTI)** Several promising techniques have been developed to elicit truthful answers from language models by probing their inner beliefs. In the appendix section I provide a short review of the Contrast-Consistent Search methodology [56] and Inference Time Intervention technique [60]. As both come with a set of limitations (see appendix B.2.1, B.3.1), I try to build on top of them and develop some (quite high-level) intuition for a novel technique, namely SSTI, that tries to combine the advantages of those approaches and unsupervised deep clustering with focal loss (see appendix B.4.2).

#### Execution Environment

**Remote Execution** The safety assessment component is executed on the servers of a trusted third-party auditor. This approach allows to establish a safety certification mechanism of R&D designs. This approach is chosen primarily for security reasons. Running the safety assessment oracle over a trusted third party ensures that the oracle’s inner workings are inaccessible to external parties, thereby reducing the attack surface. Furthermore, in terms of performances: (i) differently from the design analysis component, the oracle will likely need to be less specialised and (ii) the model of human approval embedded in the oracle is assumed to be computationally expensive.

#### Security Considerations

**Restricted Access to the Oracle** To prevent the risks associated with reverse engineering, interference, or data exploitation, access to the inner workings of the oracle is restricted.

## **Output Security**

The final output includes a digital signature from the trusted third-party, binding together the hash of the design and the safety verdict. The objective is to provide some form of certification over the safety of a certain R&D design.

### **3.4.3 Integration with the Comprehensive AI Services Model (CAIS)**

The Comprehensive AI Services model (CAIS) reframes superintelligence as a collaborative collection of specialised services rather than a single, monolithic agent. Within the CAIS framework, each service is bounded in scope and resources (but not knowledge) and performs only a specific task

## **Granularity and Component-Level Analysis**

The integration of the CAIS model with FROST-TEE unfolds a plethora of opportunities. FROST-TEE can be adapted not just to the final design but also individual units at varying granularity levels thus allowing for early testing. By deploying FROST-TEE on minor elements or modules within a design, a more in-depth and nuanced evaluation is attainable. Furthermore, it ensures enhanced scalability by using FROST-TEE over a smaller cluster of services responsible for the distinct components of a design. This method enhances performances, minimises resource cost, and identifies vulnerabilities early in the design phase.

## **Decentralisation and Delegation**

For the safety assessment of the comprehensive design, preliminary testing can be executed for individual elements and potentially be executed locally. However, for obtaining a certificate, authentication of the design's safety by an auditing trusted third party as shown above, is deemed more secure.

## **Evaluation**

The FROST-TEE system offers a robust and multi-layered approach for design analysis and safety assessment, combining AI and cybersecurity methodologies. Its strategic use of local and remote processing allows for both detailed analysis and

secure verification. However, the system is theoretically complex and incorporates unproven techniques like DLK and ITI approaches, which necessitate extensive experimentation for validation. While promising, its effectiveness and reliability are not guaranteed without rigorous testing.



# Conclusion and Future Research

This research report has provided an exploration under new lens of the AI alignment problem, focusing on the paradigm shift introduced by the Comprehensive AI Services (CAIS) model and its potential in tackling the Eliciting Latent Knowledge (ELK) problem, through FROST-TEE which aims at ensuring the safety of designs in R&D.

Still, several areas require further research and development.

One such area is the issue of collusive behaviour between services within the CAIS model. Future work should aim to formally address this point, possibly through the application of multi-party protocols between services.

Additionally, the decomposition of complex tasks into smaller tasks that can be carried out by individual bounded services may present performance issues and caveats that need to be explored further. The agentic behaviour of some of these services also warrants deeper investigation, particularly in how they interact with other services and the overall system.

The concept of fine-tuning foundation models for realising specialised services need to be expanded upon. Likewise, for the tension between decentralisation and centralisation tendencies for what concerns the CAIS infrastructure.

Further research should especially be devoted in better characterising and outlining use cases of CAIS for R&D design engineering (tackled here) and others.

As for FROST-TEE, the development of a more defined prototype, possibly implementing drastic changes to the current model, is required before any in-depth discussion can be carried out. Moreover, the techniques introduced, such as Semi-Supervised inference-time Truthful Intervention (SSTI), require necessary experimentation to validate their efficacy and reliability. Lastly, the reward signal for designer services within FROST-TEE needs to be better shaped to align with the overarching goals of safety and alignment.



# Appendix A

## Secure Design Analysis

### A.1 (Recursive) Factored Decomposition

Factored decomposition [52] is a method designed to enhance the faithfulness and accuracy of reasoning generated by large language models (LLMs) when tasked with answering questions. Specifically, factored decomposition seeks to improve the quality of this reasoning by breaking down a complex question into simpler subquestions and then synthesising an answer to the original question based on these sub-answers. This method operates in three main stages: decomposition, subquestion-answering, and recomposition.

#### A.1.1 Method

##### Stage 1: Decomposition

In the initial stage, the model is prompted with a complex question  $q$ . The model then generates a list of subquestions  $l_1 = [q_1, q_2, \dots]$  that collectively encapsulate the aspects of the original question  $q$ . Importantly, subquestions in this list  $l_1$  may contain references to answers from other subquestions, thereby allowing for interdependencies and a more accurate analysis.

##### Stage 2: Subquestion-Answering

Once  $l_1$  is established, the model begins to answer those subquestions that are independent, meaning they don't reference any other subquestions. Each subquestion

$q_i$  is posed to the model in an isolated context to generate a corresponding subanswer  $a_i$ . After obtaining the initial set of subanswers, they are collated into a list  $a = [a_1, a_2, \dots]$ .

Now, the model reviews  $l_1$  to create a new set of unanswered subquestions  $l_2$  by editing, removing, or replacing references within the subquestions in  $l_1$  based on the subanswers in  $a$ . This iterative process alternates between updating the list of subquestions and answering them until a predetermined output condition is met, indicating that the model has sufficient information to answer the original question  $q$ .

### Stage 3: Recomposition

At this point, a reasoning sample  $x$  is compiled, consisting of tuples  $(q_i, a_i)$  of each subquestion and its corresponding subanswer. This sample is then used in the final recomposition stage to answer the original question  $q$ .

#### A.1.2 Remarks

The strength of factored decomposition lies in its ability to isolate reasoning in separate contexts for each subquestion. This minimises the impact of any potential biases and unverbaised factors that might affect the answer to the original, more complex question. Moreover, by deconstructing the problem into constituent parts, the model is forced to condition its final answer more heavily on the reasoning established through the subquestions, thereby improving the faithfulness of the answer relative to methods like Chain-of-Thought (CoT).

By employing this technique *recursively* - breaking subquestions down into simpler and simpler sub-subquestions - one can achieve a high level of granularity. The ultimate aim is to reach a point where each sub-question is human-understandable, thereby enhancing oversight and explainability.

I won't go further into the details as this is meant more as a thought experiment rather than an actual system prototype. This is just to point out that techniques to decompose questions into more understandable subquestions exist and are being analysed. Other similar approaches include [61, 62, 63].

## A.2 Debate

Refer to [38, 64].

# Appendix B

## Secure Safety Oracle

Superhuman models present unique challenges in ensuring the trustworthiness of generated outputs. This appendix seeks to delve into state of the art techniques for eliciting truthful answers from an oracle, focusing on the potential of consistency checks as a strategy for improving response reliability and honesty. The underlying assumption guiding this exploration is that truthfulness correlates with consistency. By identifying logical inconsistencies, we aim to spot potential "dishonesty" in a model's responses.

The next few sections will consist of summaries of some techniques and an analysis of their limitations. I will also try to develop some intuitions for a possible novel approach that draws inspiration from the techniques under scrutiny in an attempt to address some of their shortcomings.

The purpose of this appendix is to provide a short but quite detailed literature review to show an example of how state of the art techniques can be used in the framework previously discussed. The underlying intuition is that techniques built for seemingly simple oracle models acquire new importance in the setting discussed in the paper.

### B.1 Strategy: Check Inconsistencies

The focus of this section is to explore the avenue of leveraging consistency checks to assess the trustworthiness of language models. This method involves analysing the model's responses across a range of queries to identify any inconsistencies. Importantly, this approach may provide a more empirical foundation for tackling the

ELK problem.

The rationale behind this strategy is straightforward. We can think of a model that tries to produce answers that are most convincing to humans for each specific query. However, if we take into account the relationships between different queries, the model would need to consider multiple potential outputs before arriving at a convincing enough answer. The objective is to reach a scenario in which providing truthful and consistent answers becomes the most computationally efficient strategy. As noted in the discussion of the CAIS model, optimisation pressure can bring safety. In particular, we could hope that such a pressure leads the model to track consistency and possibly truthfulness among its hidden features.

Yet, this is not a foolproof solution. A sophisticated model might still find a way to appear consistent while not being truthful, tailoring its answers to pass human-conducted consistency checks. The open question, then, is whether a faking consistency strategy could be easily found and executed by an oracle. If the answer is affirmative, then the approach may have a fundamental flaw. On the other hand if sufficiently hard consistency checks can be found, this may yield computational advantages that make deceitful strategies untenable. The hope is that in the limit, being truthful is the most efficient strategy for the model instead of it keeping track of all consistency checks human operators will apply.

### B.1.1 Methodology for Evaluating Superhuman Models

Let's explore a methodology for evaluating machine learning models that may have capabilities exceeding human expertise. Traditional evaluation benchmarks often fall short because they rely on human judgement, which may not serve as an adequate baseline for these advanced models. As argued above, among various proposed solutions, consistency checks stand out as a viable option for detecting dishonest or unreliable outputs. The following discussion is informed by [65].

Consider a model  $f$ . Our objective is to identify sets of queries  $x_1, \dots, x_n$  for which the model's outputs may violate established consistency rules. Although verifying the absolute accuracy of individual outputs  $y'_1, y'_2, \dots$  may not be feasible, any inconsistency among them would suggest that at least one of the outputs is incorrect.

The limitation in this kind of approaches is that while it can be used to show model inconsistencies, not finding any does guarantee their absence. Nonetheless,

consistency checks can provide a useful way for building contrast pairs of inputs on which to test a model, which is the main intuition that carries over to the next part of the discussion.

## B.2 Inference Time Intervention

The Inference-Time Intervention (ITI) technique [60] aims to improve the accuracy of large language models by aligning what the model "knows" with what it actually "tells" the user. ITI works by adjusting the model's internal activations during the inference stage. Specifically, it shifts activations along an identified "truthfulness direction" for a selected subset of the model's attention heads, thus guiding the output towards greater truthfulness. The discussion below follows that in [60].

### Transformer-based Model

Consider a transformer-based model:

$$x_{l+1} = x_l + \sum_{h=1}^H Q_l^h \text{Att}_l^h(P_l^h x_l)$$

where ( $P$  is the mapping from activation space to head space and  $Q$  is the inverse mapping.

### Probing Dataset

Given  $\{q_i, a_i, y_i\}$  from a labelled dataset, concatenate  $q_i \parallel a_i$  and take out head activations at the last token to collect the probing dataset  $\{(x_i^h)_l, y_i\}_{i=1}^N$  for every head  $h$ , for every layer  $l$  (i.e. we have a probe for each head and layer).

### Probe Training

Train the probe [66] with:

$$p_\theta(x_l^h) = \text{sigmoid}(\langle \theta, x_l^h \rangle)$$



Notice that by training  $p'_\theta$  with constraint  $\theta \perp \theta'$ , we obtain better than chance accuracy, indicating that truth lies in a subspace rather than in a single direction.

In this regard, future research avenues could explore the use of multi-dimensional probing techniques (e.g. by manifold learning) to more accurately capture the complex geometry where truth resides. This could lead to interventions that are not only more effective but also more nuanced in enhancing a model's truthfulness. Furthermore, the notion of a truth subspace opens up the possibility of employing ensemble methods that combine multiple truthful directions.

### Inference Time Intervention

Once the probe is trained, perform intervention at inference time:

- From training, select only the top  $K$  heads most related to truth (minimally invasive approach). Rank truth-relatedness of attention heads by measuring probe accuracy for head  $h$  on the validation set.
- Determine the vector along which to shift activations in head output. Here the authors consider two options:
  1. Take vector perpendicular to the hyperplane learned by the probe.
  2. Take vector connecting true and false distributions. For every layer  $l$  in layer set, for every head  $h$  in head set:
    - (a) Compute the average of  $(x^h)_l$  values for **true** and **false** labels respectively:

$$\mu_{\text{true}_l^h} = \frac{1}{N_{\text{true}}} \sum_{i=1}^{N_{\text{true}}} (x_{l,i}^h)_{\text{true}} \quad (\text{B.1})$$

$$\mu_{\text{false}_l^h} = \frac{1}{N_{\text{false}}} \sum_{j=1}^{N_{\text{false}}} (x_{l,j}^h)_{\text{false}} \quad (\text{B.2})$$

- (b) Compute the centre of mass direction  $\psi$  for each head and layer:

$$\psi_l^h = \mu_{\text{true}_l^h} - \mu_{\text{false}_l^h} \quad (\text{B.3})$$

- Shift activation along the truthful direction by modifying MHA:

$$x_{l+1} = x_l + \sum_{h=1}^H Q_l^h (\text{Att}_l^h(P_l^h x_l) + \alpha \sigma_l^h \psi_l^h). \quad (\text{B.4})$$

where  $\psi$  is the truthful direction (it is a zero vector for non-selected heads),  $\alpha$  is the shifting coefficient, and  $\sigma$  is the standard deviation of activations along truthful directions.

- Repeat the procedure for each next token prediction in an auto-regressive manner until the oracle answer is generated

### B.2.1 Supervised vs. Unsupervised

The Inference-Time Intervention (ITI) technique relies on supervised learning to identify directions in the model’s activation space that are associated with truthfulness. However, this approach raises concerns about the validity of the learned representations. Specifically, the probes may not necessarily capture the actual representation of truth, but rather latch onto features that are spuriously correlated with truthfulness in the training data [67]. The problem is that such correlations may not generalise well to more general contexts or unseen data.

These issues stem from the inherent limitations of supervised learning models and how they handle unknown scenarios and unseen data patterns. The shortcomings of supervised learning are further exacerbated from the fact that when dealing with superhuman model it is hard to come up with groundtruths in the first place. In the following section, we will explore an alternative approach that attempts to address the problem of eliciting truthfulness in an unsupervised manner.

## B.3 Discovering Latent Knowledge

The second approach for training a probe does not require labelled data. Instead, it attempts to identify patterns in the language model’s embeddings that satisfy certain logical coherence properties. Here we see how consistency checks may be useful. An implementation of this idea is the Contrast-Consistent Search (CCS) technique developed in [56]. I outline the key points of the discussion below.

## Problem

Language models’ training techniques can be misaligned with the truth, causing the model to output falsehoods even if it may know better.

- Imitation learning may induce the model to reproduce errors humans make.
- Training a model to produce text that humans rate highly may be prone to errors that human operators cannot evaluate.
- Chatbot trained to optimise engagement may output text that is compelling but false.

## Solution

We can tackle these issues by extracting hidden knowledge from a language model’s internal activations using an entirely unsupervised approach. Specifically, [56] offers a method to accurately respond to yes-no questions using unlabelled model activations. It accomplishes this by identifying a direction in the activation space that upholds logical consistency.

Notice that while in the context of the paper accurately answer yes-no question may resemble a toy scenario, in our setting, this is exactly the kind of capability we are looking for. This serves as evidence that techniques developed for simpler scenarios, under the CAIS model could be effective even when applied to more complex use cases. In contrast, with a monolithic agent model, it’s less obvious how findings from [56] could be applied to more intricate situations.

This advancement enhances our understanding of what language models actually know, as opposed to what they say. For instance, the method maintains high accuracy even when the models are prompted to give incorrect answers. Additionally, the technique does not require access to explicit ground truth labels, a particularly important feature in the context of superhuman models, where the ground truth may not always be accessible.

### B.3.1 CCS Methodology

Consider feature representation  $\phi(x)$  on a natural language input  $x$ . Our goal is to answer the questions  $q_1, \dots, q_n$  only given access to  $\phi(\cdot)$ . Hence, we only consider hidden states.

### Step 1: Constructing Contrast Pairs

For each question  $q_i$ , the model answers it in two ways: once with "Yes" and once with "No". This results in two new natural language statements  $x_i^+$  and  $x_i^-$ . The labels used for constructing contrast pairs can vary depending on the task. In the context of FROST-TEE, we may consider the design analysis coming from factored decomposition

**Example:**

$x^+ = \text{"Q: Is the design D given property P safe or unsafe? A: safe"}$

$x^- = \text{"Q: Is the design D given property P safe or unsafe? A: unsafe"}$

### Step 2: Feature Extraction and Normalisation

Compute the representation of  $\phi(x_i^+)$  and  $\phi(x_i^-)$ . Normalise the representations as shown below. The objective is to isolate the dimension of truthfulness in the hidden representations  $\phi(x_i^+)$  and  $\phi(x_i^-)$ . That is, the aim is to neutralise the influence of the differing terminal tokens thereby ensuring that any difference is supposedly solely attributable to the truthfulness of the statements.

$$\tilde{\phi}(x_i^+) = \frac{\phi(x_i^+) - \mu^+}{\sigma^+}$$

$$\tilde{\phi}(x_i^-) = \frac{\phi(x_i^-) - \mu^-}{\sigma^-}$$

Here,  $\mu^+$  and  $\sigma^+$  are the mean and standard deviation of  $\{\phi(x_i^+)\}_{i=1}^n$ , and  $\mu^-$  and  $\sigma^-$  are the mean and standard deviation of  $\{\phi(x_i^-)\}_{i=1}^n$ . All operations are element-wise along each dimension.

### Step 3: Mapping Activations to Probabilities

Train a linear probe [66] to classify  $x_i^+$  and  $x_i^-$  as true or false. That is, the method learns a probe  $p_{\theta,b}(\tilde{\phi})$  that maps a normalised hidden state  $\tilde{\phi}$  to a value between 0 and 1 interpreted as the probability of a statement to be true:

$$p_{\theta,b}(\tilde{\phi}) = \sigma(\theta^T \tilde{\phi} + b)$$

Notice that the use of more advanced probes could also be investigated [68].

The objective is to optimise the mapping such that probabilities are both consistent and confident. The loss function  $L$  can be represented as:

$$L_{\text{consistency}}(\theta, b; q_i) := \left[ p_{\theta,b}(\tilde{\phi}(x_i^+)) - (1 - p_{\theta,b}(\tilde{\phi}(x_i^-))) \right]^2$$

$$L_{\text{confidence}}(\theta, b; q_i) := \min\{p_{\theta,b}(\tilde{\phi}(x_i^+)), p_{\theta,b}(\tilde{\phi}(x_i^-))\}^2$$

$$L_{\text{total}}(\theta, b) := \frac{1}{n} \sum_{i=1}^n L_{\text{consistency}}(\theta, b; q_i) + L_{\text{confidence}}(\theta, b; q_i)$$

#### Step 4: Inference

Once the probe has been trained, we can use it for new questions. For the sake of simplicity, we can consider the answer to  $q_i$  to be “Yes” if  $p_{\text{avg}} > 0.5$  where:

$$p_{\text{avg}}(q_i) := \frac{1}{2}(p(\tilde{\phi}(x_i^+)) + (1 - p(\tilde{\phi}(x_i^-))))$$

Refer to the original paper [56] for more details.

#### Finding Truth Features

CCS identifies representations of truth that are independent of the task at hand. Its ability to generalise across disparate tasks while maintaining competitive transfer accuracy suggests that the method can pinpoint salient features related to truthfulness, irrespectively of the task. This hints at the possibility of having oracles performing the safety assessment for more than one type of design analysis (while on the other hand design analysis step probably benefits from some fine tuning on

the specific task, emphasising the advantages of executing design analysis locally and safety assessment remotely).

## Limitations

CCS operates under the assumption that there exists a direction in the activation space that well distinguishes between true and false inputs. Specifically, it presupposes that a supervised probe could achieve high accuracy if it had access to ground truth labels. This raises questions about the conditions under which a model can both evaluate the truthfulness of an input and actively do so. The exact conditions for this remain ambiguous.

Several other limitations have been highlighted in the literature [67, 69]:

- **Caveat on Probabilistic Coherence:** CCS employs two loss functions:  $L_{\text{consistency}}$  and  $L_{\text{confidence}}$ . These functions do not account for the actual model’s accuracy but rather penalise the probe for lack of confidence and probabilistic incoherence. In other words, while the CCS probe aims to achieve probabilistic coherence, its outputs do not necessarily reflect the model’s subjective probabilities. The  $L_{\text{confidence}}$  loss function encourages the probe to report extreme values close to 0 or 1, irrespective of the model’s actual beliefs.
- **Accuracy Observations:** Further experimental results indicate that the accuracy of CCS is often no better than random chance. The probes appear to achieve low loss by identifying embeddings corresponding to sentences with negations, despite the normalisation process. This suggests that the current normalisation technique is insufficient for masking the grammatical structure of the sentences. A more sophisticated normalisation could probably work better.
- **Normalisation and Class Imbalance:** The normalisation process has been shown to underperform in the presence of class imbalance. Mean normalisation ensures that the average component along the truthful direction for the normalised representations of  $x_i^+$  and  $x_i^-$  is zero. This becomes problematic when applied to an imbalanced dataset, as it erroneously removes the average truth component from  $x_i^+$  and  $x_i^-$  representations.

### B.3.2 An alternative: Contrastive Representation Clustering (CRC)

We have discussed the Contrast-Consistent Search (CCS) technique and its limitations. To address these shortcomings I will try to broaden the scope of the investigation by also turning the attention to the Contrastive Representation Clustering (CRC) method [56]. CRC, like CCS, aims to find salient features in the representation space that are correlated with the truth. However, it employs a different approach, focusing on clustering techniques. By examining CRC, the aim is to draw inspiration from its clustering-based methodology, which could offer alternative or complementary strategies for truth evaluation.

This section outlines the procedure for CRC using Bimodal Salience Search (BSS), as shown in [56].

#### Step 1: Creation of Contrast Pairs

- **Contrast Pairs:** Form pairs  $(x_i^+, x_i^-)$  where one is true, and the other is false. The activations are represented by  $\phi(x_i^+)$  and  $\phi(x_i^-)$ .

#### Step 2: Normalisation

- **Calculate Mean and Standard Deviation:** Compute the means and standard deviations for positive and negative representations:  $(\mu^+, \sigma^+)$ ,  $(\mu^-, \sigma^-)$ .
- **Normalise Representations:** Apply normalisation to the contrast pair activations:

$$\phi^\sim(x_i^+) := \frac{\phi(x_i^+) - \mu^+}{\sigma^+}; \quad \phi^\sim(x_i^-) := \frac{\phi(x_i^-) - \mu^-}{\sigma^-}$$

#### Step 3: Bimodal Salience Search (BSS)

- **Construct Contrast Features:** Form the contrast features:

$$c_i := \phi^\sim(x_i^+) - \phi^\sim(x_i^-)$$

- **Clustering Loss:** The loss function for clustering is:

$$L(\theta) = \frac{\text{var}\{\theta^T c_i | \theta^T c_i < 0\} + \text{var}\{\theta^T c_i | \theta^T c_i \geq 0\}}{\text{var}\{\theta^T c_i\}}$$

where  $\theta$  is the direction in the representation space.

#### Step 4: Clustering

- **Project:** Project the normalised contrast pair activations onto the direction:

$$\theta^T c_i$$

- **Thresholding:** Cluster by thresholding at 0:

$$\text{Cluster 1} = \{\theta^T c_i < 0\}, \text{ Cluster 2} = \{\theta^T c_i \geq 0\}$$

#### Remarks

While experiments in [56] show that CCS seem to perform best, CRC still obtains high accuracy and achieve competitive results with respect to zero-shot baseline.

## B.4 Semi-Supervised Truthful Intervention

In an attempt to address the limitations above, I develop some intuitions for a novel method, Semi-Supervised Truthful Intervention (SSTI), for eliciting truthful answers in a language model. It builds on the respective advantages of CCS (and CRC) [56] and ITI [60] and integrates some insights from unsupervised deep clustering [70].

### B.4.1 Unsupervised Deep Clustering and Focal Loss

The work in [70] employs focal loss penalty term within a deep clustering framework to enhance the label assignment mechanism. Focal loss is traditionally used in supervised learning to handle class imbalance by focusing on hard-to-classify examples. The challenge lies in adapting focal loss for unsupervised scenarios, as it typically requires true labels for its computation. The solution proposed in [70] is



to use the target distribution in the clustering layer as a surrogate for true labels, thereby enabling the application of focal loss in an unsupervised setting.

### Clustering Loss

The clustering loss  $l_{cl}$  is defined using the KL-divergence between the soft labels distribution  $S$  and the target distribution  $T$ :

$$l_{cl} = KL(T||S) = \sum_i \sum_j t_{ij} \log \frac{t_{ij}}{s_{ij}}.$$

The soft labels distribution  $S$  is calculated using Student's  $t$ -distribution as follows:

$$s_{ij} = \frac{(1 + ||h_i - \mu_j||^2/\alpha)^{-\frac{\alpha+1}{2}}}{\sum_j (1 + ||h_i - \mu_j||^2/\alpha)^{-\frac{\alpha+1}{2}}}$$

where  $\alpha$  is fixed to 1 and controls the extent of freedom of Student's  $t$ -distribution. The target distribution  $T$  is then formalised as:

$$t_{ij} = \frac{s_{ij}^2 / \sum_i s_{ij}}{\sum_j s_{ij}^2 / \sum_i s_{ij}}$$

### Focal Loss

For binary classification,  $p$  and  $y$  denote the predicted and true labels, respectively. In this context, true labels are generally required, but this is inconsistent with the unsupervised nature of the clustering task we are considering. However, as anticipated earlier, the soft labels distribution  $S$  and target distribution  $T$  may offer a solution. Specifically, the predicted labels are generated from  $S$  and the true labels are generated from  $T$ . The focal loss  $l_{fl}$  can thus be defined as:

$$l_{fl} = -(1 - p_t)^\gamma \log(p_t),$$

where

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise} \end{cases}$$

and  $\gamma$  is a tunable parameter that satisfies  $\gamma \geq 0$ .

In other words,  $-(1 - p_t)^\gamma$  is the modulating factor of  $\text{CEloss}(p_t) = -\log(p_t)$

Focal loss is incorporated to enhance label assignment by focusing on hard-to-classify samples. Specifically,  $l_{fl}$  increases the loss contribution from hard samples where  $p_t$  is close to 0, while diminishing the impact of easy samples where  $p_t$  is close to 1.

### Combined Objective

The combined objective function  $L_c$  for the clustering module is:

$$L_c = l_{cl} + l_{fl},$$

where  $l_{cl}$  enables self-training-based clustering, and  $l_{fl}$  addresses the issue of hard-to-classify samples.

### Updating Cluster Centroids

During training, the cluster centroids  $\mu_j$  are updated employing stochastic gradient descent (SGD) and backpropagation. With  $\rho$  as the learning rate and  $m$  as the mini-batch size, the update formula for  $\mu_j$  is:

$$\mu_j = \mu_j - \frac{\rho}{m} \sum_{i=1}^m \frac{\partial L_c}{\partial \mu_j}$$

## B.4.2 SSTI Methodology

This work aims to incorporate these insights into the SSTI method, particularly the use of focal loss for handling hard samples, to improve the elicitation of truthful responses from language models.

### Step 1: Supervised Start

Here I assume the availability of a labelled dataset for a supervised start (something similar to TruthfulQA [71] benchmark — I won't delve further into the matter).

- **Training and Ranking Heads:** Train a probe over the head values to identify the truth-relatedness of attention heads, similar to the ITI technique. Rank the first  $K$  heads by probe accuracy over the validation set.
- **Initialise Centroids for Each Head and Layer:** Use the labelled dataset to initialise the centroids for K-divergence loss with respect to head values for true and false statements. These centroids, denoted as  $\mu_{\text{true}_l^h}$  and  $\mu_{\text{false}_l^h}$ , are initialised for each layer  $l$  and each of the top  $K$  heads  $h$ .

## Step 2: Creation of Contrast Pairs

### 1. Identification of Key Properties:

From the design analysis, identify key properties important for safety assessment. For the routing algorithm discussed in section 2.6.1, these could include:

- Maximum latency of 50 ms
- 99.999% up-time
- Real-time adaptability
- Continuous access to all data packets

### 2. Formulation of Questions:

Create questions that encapsulate these key properties in the context of safety. For instance:

- “Is the design of {Routing Algorithm} with {Maximum latency of 50 ms} safe or unsafe?”
- “Is the design of {Routing Algorithm} with {Continuous access to all data packets} safe or unsafe?”

### 3. Creation of Contrast Pairs:

For each question, create a pair of statements that answer it in the affirmative and the negative. These will serve as your  $x_i^+$  and  $x_i^-$  pairs. For instance:

- $x_i^+$ : “Q: Is the design of {Routing Algorithm} with {Maximum latency of 50 ms} safe or unsafe? A: safe”
- $x_i^-$ : “Q: Is the design of {Routing Algorithm} with {Maximum latency of 50 ms} safe or unsafe? A: unsafe”

### Step 3: Unsupervised Clustering with Focal Loss

- **Iterate Over Contrast Pairs:** Iterate over the contrast pairs, clustering the heads of the transformer-based language model around the two centroids  $\mu_{\text{true}_l^h}$  and  $\mu_{\text{false}_l^h}$ .
- **Clustering Loss and Focal Loss Integration:** Refer to section B.4.1.

$$L_c = l_{cl} + l_{fl}$$

- **Update Centroids:** Update the centroids  $\mu_j$  using stochastic gradient descent (SGD) and backpropagation:

$$\mu_j = \mu_j - \frac{\rho}{m} \sum_{i=1}^m \frac{\partial L_c}{\partial \mu_j}$$

### Step 4: Compute Truthful Direction for Each Layer and Head

- **Determine the Vector for Shifting Activations:** Compute the direction  $\psi$  for each head  $h$  and layer  $l$ :

$$\psi_l^h = \mu_{\text{true}_l^h} - \mu_{\text{false}_l^h}$$

### Step 5: Inference Time Intervention

- **Shift Activation Along the Truthful Direction:** Modify the Multi-Head Attention (MHA) to shift activations along the truthful direction  $\psi$ :

$$x_{l+1} = x_l + \sum_{h=1}^H Q_l^h (\text{Att}_l^h(P_l^h x_l) + \alpha \sigma_l^h \psi_l^h)$$

where  $\psi$  is the truthful direction (it becomes a zero vector for non-selected heads),  $\alpha$  is the shifting coefficient, and  $\sigma$  is the standard deviation of activations along truthful directions.

### B.4.3 Conclusion

#### Experimental Evaluation

While the SSTI methodology within the FROST-TEE framework tries to address the shortcomings of state of the art techniques in eliciting truthful responses from transformer-based language models, it is essential to note that the methodology above is currently sketched at a high level only. Formal experiments are crucial to validate the effectiveness and robustness of the approach and correct any mistake. Specifically, empirical studies should be conducted to evaluate the accuracy of the probe, the benefits (if any) of a semi-supervised approach, the impact of the K-divergence and focal loss on clustering, and the efficacy of the inference-time intervention. Additionally, the initialisation and updating of centroids for each layer and head need to be empirically validated to ensure they contribute meaningfully to the truthfulness of the model’s outputs.

#### Future Research Directions

The SSTI method may offer an approach to elicit truthful answers from language models by combining supervised and unsupervised techniques. Future research could investigate the introduction of contractive features to for improving robustness and learn more discriminative representation of truth.

# Bibliography

- [1] Evan Hubinger. An overview of 11 proposals for building safe advanced ai, 2020.
- [2] Paul Christiano. Prosaic AI alignment — medium, March 2017.
- [3] Iason Gabriel. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3):411–437, September 2020.
- [4] Stuart Russell. Human compatible. ai and the problem of control, london: Allen lane, 2019.
- [5] Dylan Hadfield-Menell and Gillian K. Hadfield. Incomplete contracting and ai alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’19, pages 417–422, New York, NY, USA, 2019. Association for Computing Machinery.
- [6] Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.
- [7] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- [8] Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML ’04, page 1, New York, NY, USA, 2004. Association for Computing Machinery.
- [9] Markus Peschl, Arkady Zgonnikov, Frans A. Oliehoek, and Luciano C. Siebert. Moral: Aligning ai with human norms through multi-objective reinforced active learning. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS ’22, page 1038–1046, Richland, SC, 2022. International Foundation for Autonomous Agents and Multiagent Systems.
- [10] Dizan Vasquez, Billy Okal, and Kai Arras. Inverse reinforcement learning algorithms and features for robot navigation in crowds: an experimental comparison. *IEEE International Conference on Intelligent Robots and Systems*, 10 2014.
- [11] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values, 2023.
- [12] Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning, 2017.

- 
- [13] Alasdair MacIntyre. *After virtue*. A&C Black, 2013.
  - [14] John McDowell. Virtue and reason. *The monist*, 62(3):331–350, 1979.
  - [15] Shannon Vallor. *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press, 2016.
  - [16] DeepMind Research. Specification gaming: the flip side of AI ingenuity.
  - [17] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Inc., USA, 1st edition, 2014.
  - [18] K.E. Drexler. Reframing superintelligence: Comprehensive ai services as general intelligence. Technical Report 2019-1, Future of Humanity Institute, University of Oxford, 2019.
  - [19] Alexander Matt Turner, Logan Riggs Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. Optimal policies tend to seek power. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.
  - [20] Alexander Matt Turner and Prasad Tadepalli. Parametrically retargetable decision-makers tend to seek power. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.
  - [21] Evan Hubinger. How likely is deceptive alignment? — LessWrong.
  - [22] DeepMind Safety Research. Goal Misgeneralisation: Why Correct Specifications Aren’t Enough For Correct Goals | by DeepMind Safety Research | Medium.
  - [23] George E Monahan. State of the art—a survey of partially observable markov decision processes: theory, models, and algorithms. *Management science*, 28(1):1–16, 1982.
  - [24] Embedded Agency (full-text version) — LessWrong.
  - [25] R.J. Solomonoff. A formal theory of inductive inference. part i. *Information and Control*, 7(1):1–22, 1964.
  - [26] Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. Truthful ai: Developing and governing ai that does not lie, 2021.
  - [27] H. Akin Ünver. Artificial intelligence, authoritarianism and the future of political systems. Technical report, Centre for Economics and Foreign Policy Studies, 2018.
  - [28] Planning for AGI and beyond.
  - [29] Lin Padgham and Michael Winikoff. *Developing intelligent agent systems: A practical guide*. John Wiley & Sons, 2005.
  - [30] John von Neumann, Oskar Morgenstern, and Ariel Rubinstein. *Theory of Games and Economic Behavior (60th Anniversary Commemorative Edition)*. Princeton University Press, 1944.

- [31] J. Barrat. *Our Final Invention: Artificial Intelligence and the End of the Human Era*. St. Martin's Press, 2013.
- [32] N Bostrom. Existential risks: analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology*, 9, 2002.
- [33] Mikhail Batin, Alexey Turchin, Markov Sergey, Alisa Zhila, and David Denkenberger. Artificial intelligence in life extension: from deep learning to superintelligence. *Informatica*, 41(4), 2017.
- [34] Anthony M. Barrett and Seth D. Baum. A model of pathways to artificial superintelligence catastrophe for risk and decision analysis. *Journal of Experimental; Theoretical Artificial Intelligence*, 29(2):397–414, may 2016.
- [35] Irving John Good. Speculations concerning the first ultraintelligent machine\*\*based on talks given in a conference on the conceptual aspects of biocommunications, neuropsychiatric institute, university of california, los angeles, october 1962; and in the artificial intelligence sessions of the winter general meetings of the iee, january 1963 [1, 46].the first draft of this monograph was completed in april 1963, and the present slightly amended version in may 1964.i am much indebted to mrs. euthie anthony of ida for the arduous task of typing. volume 6 of *Advances in Computers*, pages 31–88. Elsevier, 1966.
- [36] Michael Austin Langford and Betty H.C. Cheng. A modular and composable approach to develop trusted artificial intelligence. In *2022 IEEE International Conference on Autonomic Computing and Self-Organizing Systems (ACSOS)*, pages 121–130, 2022.
- [37] Mark d’Inverno, Michael Luck, Michael Georgeff, David Kinny, and Michael Wooldridge. The dmars architecture: A specification of the distributed multi-agent reasoning system. *Autonomous Agents and Multi-Agent Systems*, 9:5–53, 07 2004.
- [38] Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate, 2018.
- [39] K.E. Drexler. Mdl intelligence distillation: Exploring strategies for safe access to superintelligent problem-solving capabilities. Technical Report 2015-3, Future of Humanity Institute, Oxford University, 2015.
- [40] Matthew Barnett. Updating Drexler’s CAIS model.
- [41] Eric Drexler. “Reframing Superintelligence” + LLMs + 4 years.
- [42] Reframing Superintelligence: Comprehensive AI Services as General Intelligence — Less-Wrong.
- [43] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.
- [44] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks, 2013.



- [45] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization, 2015.
- [46] David Ferraiolo, Janet Cugini, D Richard Kuhn, et al. Role-based access control (rbac): Features and motivations. In *Proceedings of 11th annual computer security application conference*, pages 241–48, 1995.
- [47] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, oct 2016.
- [48] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS '17*, page 1175–1191, New York, NY, USA, 2017. Association for Computing Machinery.
- [49] Payman Mohassel and Peter Rindal. Aby3: A mixed protocol framework for machine learning. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS '18*, page 35–52, New York, NY, USA, 2018. Association for Computing Machinery.
- [50] Chiraag Juvekar, Vinod Vaikuntanathan, and Anantha Chandrakasan. GAZELLE: A low latency framework for secure neural network inference. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 1651–1669, Baltimore, MD, August 2018. USENIX Association.
- [51] Jessica Taylor. Quantilizers: A safer alternative to maximizers for limited optimization. In *AAAI Workshop: AI, Ethics, and Society*, 2016.
- [52] Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiuūtė, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Sam McCandlish, Sheer El Showk, Tamara Lanham, Tim Maxwell, Venkatesa Chandrasekaran, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Question decomposition improves the faithfulness of model-generated reasoning, 2023.
- [53] Christof Ferreira Torres, Antonio Ken Iannillo, Arthur Gervais, and Radu State. Confuzzius: A data dependency-aware hybrid fuzzer for smart contracts. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 103–119, 2021.
- [54] Xiaojun Chen, Shengbin Jia, and Yang Xiang. A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications*, 141:112948, 2020.
- [55] Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkovich, Ranit Aharonov, and Noam Slonim. Are you convinced? choosing the more convincing evidence with a siamese network, 2019.

- [56] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision, 2022.
- [57] Paul Christiano, Ajeya Cotra, and Mark Xu. Eliciting Latent Knowledge.
- [58] Marius Hobbhahn. Eliciting Latent Knowledge (ELK) - Distillation/Summary – alignment-forum.
- [59] George Coker, Joshua Guttman, Peter Loscocco, Amy Herzog, Jonathan Millen, Brian O’Hanlon, John Ramsdell, Ariel Segall, Justin Sheehy, and Brian Sniffen. Principles of remote attestation. *International Journal of Information Security*, 10(2):63–81, 2011.
- [60] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model, 2023.
- [61] Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models, 2023.
- [62] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models, 2023.
- [63] Antonia Creswell, Murray Shanahan, and Irina Higgins. Selection-inference: Exploiting large language models for interpretable logical reasoning, 2022.
- [64] Vojtěch Kovařík and Ryan Carey. (when) is truth-telling favored in ai debate?, 2021.
- [65] Lukas Fluri, Daniel Paleka, and Florian Tramèr. Evaluating superhuman models with consistency checks, 2023.
- [66] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2018.
- [67] B. A. Levinstein and Daniel A. Herrmann. Still no lie detector for language models: Probing empirical and conceptual roadblocks, 2023.
- [68] Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances, 2021.
- [69] Tom Angsten and Ami Hays. Ground-Truth Label Imbalance Impairs the Performance of Contrast-Consistent Search (and Other Contrast-Pair-Based Unsupervised Methods) — AlignmentForum.
- [70] Jinyu Cai, Shiping Wang, Chaoyang Xu, and Wenzhong Guo. Unsupervised deep clustering via contractive feature representation and focal loss. *Pattern Recognition*, 123:108386, 2022.
- [71] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022.



Eidgenössische Technische Hochschule Zürich  
Swiss Federal Institute of Technology Zurich

**Title of work:**

# Eliciting Latent Knowledge in Comprehensive AI Services Models

A Conceptual Framework and Preliminary Proposals for AI Alignment and Safety in R&D

**Thesis type and date:**

Research Project, July 14, 2023

**Supervision:**

Patrick Levermore

**Student:**

Name: Alessandro Cabodi  
E-mail: acabodi@student.ethz.ch  
Legi-Nr.:

**Statement regarding plagiarism:**

By signing this statement, I affirm that I have read and signed the Declaration of Originality, independently produced this paper, and adhered to the general practice of source citation in this subject-area.

Declaration of Originality:

[http://www.ethz.ch/faculty/exams/plagiarism/confirmation\\_en.pdf](http://www.ethz.ch/faculty/exams/plagiarism/confirmation_en.pdf)

Zurich, 13. 11. 2023:

*My Signature*