Toward A Normative Account of Machine Learning Explanation Via Levels of Abstraction

Kuil Schoneveld*

October 9, 2023

Abstract

The aim of this thesis will be to bridge between the domains of explainable artificial intelligence (XAI) and the normative criteria required for the responsible deployment of such models. The innate difficulty in understanding complex information processing systems such as those constituting the field of artificial intelligence motivates the need for methods to untangle their inner workings. Toward this end, I argue for the use of a fundamental epistemological method - that of Levels of Abstraction (LoAs) - for clarifying the workings of such systems.

I begin by articulating a predominant account of scientific understanding from Kareem Khalifa to argue that opacity, as the main obstacle to understanding, is a phenomenon relative to those seeking an explanation (Section 2). After describing the Method of LoAs, I motivate a transition from using Marr's levels of analysis to LoAs in the domain of AI to ground normative criteria for comparing explanations (Section 3). I then provide further examples of the usefulness of LoAs in the domain of AI for the sake of conceptualizing the responsibility gap and understanding advanced properties in AI models (Section 4).

1 Introduction

This introductory section will briefly elaborate on some of the background knowledge helpful in making sense of how explanation and understanding are achieved in AI. This overviews the nature of our normative target and sketches our path toward it. Here, we will see much of the initial literature review, despite some of the material no longer remaining centrally relevant[1]. Nonetheless,

*7129019 - k.g.schoneveld@students.uu.nl

[1]Despite this, I believe it is worth including this review of the responsibility gap literature for the sake of situating the end goal of normative accounts of explanation in AI. More specifically, although the primary contribution of this thesis is directed toward making normative distinctions between various stakeholder's explanatory requirements, discussing responsibility gaps represents a further relevant next step in the research and is mentioned in Section 4.1.

the literature review helps to situate the primary aims established at the beginning of this project. These aims are then expounded upon in the form of the research questions which guided this work from its outset.

Following this, Section 2 gives an account of understanding which is relevant for fulfilling the explanatory requirements of various stakeholders who interact with- and are affected by advanced AI models. After describing the components furnishing a central account of understanding, I provide an overview of a taxonomy of opacity in AI. Next, connecting this account to understanding to the taxonomy of opacity begins to reveal how the blockages to our understanding may be overcome.

Next, Section 3 articulates the Levels of Abstraction (LoA) methodology, which I argue to be the main tool for overcoming opacity. I begin by describing the method of LoAs in detail, and showing how similar frameworks have already been used to analyze algorithms in AI. I then argue that one commonly-used similar framework is a weaker version of this LoA approach and that, in fact, the same process can be accomplished with more precision using LoAs. To complete this section, I attempt to integrate the account of understanding as described in Section 2 with these LoAs to offer novel methods for explanation.

Finally, Section 4 briefly indicates some avenues for further research along the lines of the methodology described. In particular, I note an inconsistency within an argument against the existence of responsibility gaps, outline the direction for a potential research program involving LoA in the clarification of normative criteria, and mention the use of LoAs in understanding advanced AI capabilities.

1.1 Background Note on LoA Method

Since it was not present in the initial literature scan, a placeholder definition of a LoA is worth including here, the reflections of which will be recognized throughout the thesis leading into its full definition in Section 3. Specifically, LoAs are frameworks for analyzing systems based on their input and output variables to understand their function while all else is abstracted away. I offer a more comprehensive definition of these levels later. For now however, we can interpret thought through neural, psychological, or social levels. A neural level includes the examining of input and output variables in terms of electrochemical signals within synapses, as well as the neural functionality transmitting spike train patterns across brain regions. Nonetheless, this approach requires the abstraction of substantial information that may appear relevant, such as the impact of propaganda (social) or individual desires (psychological). This abstraction serves as a distinguishing characteristic, setting a given LoA apart from alternatives.

1.2 Literature Review

The following section outlines some of the preliminary reading related to the topic of this thesis. Although the notion of responsibility gaps admittedly plays

less of a role itself as a target for this thesis, it is important to describe such that we may usefully place ourselves within the domain. More precisely, responsibility gaps represent the amorphous challenge taken up by normative accounts of explainable AI. As such, I begin by explaining some of the background behind responsibility gaps to make sense of what various accounts of understanding in AI are seeking (Section 1.2.1). I then overview some of the explanation techniques involved in attempting to address the concerns born out of the responsibility gap discussion (Section 1.2.2). I then describe some work in the area of understanding and its relationship with the complexity of AI models (Section 1.2.3).

1.2.1 Responsibility Gaps

Much has been written about the potential responsibility gaps in artificial intelligence (AI) and various formulations have been proposed. The earliest formulation can be attributed to Andreas Matthias, who is one of the first to argue that there will be "an increasing class of machine actions incompatible with traditional means of responsibility ascription because nobody has meaningful control" [36]. This statement ties a thread between the increasing autonomy and decreasing clarity in responsibility attribution. It is this thread of increasing uncertainty between a clear - and in some ways intended - loss of control over new autonomous systems that dissolves into a frayed end which does not link well with our current moral and legal frameworks that is shared by all accounts of the responsibility gap. As programmers lose control over their automata, as their range of possible actions widens, and the inner workings are obscured by increasing opacity from their complex design, we find that these useful machines begin to slip further from our grasp.

Along an immediate tangible-to-abstract dimension, some expansions on this idea have identified case studies where automated decision-making directly leads to human harm [54], or more conceptual issues associated with the fuller scope of harms made possible to humans [11]. Others have striven to keep the human side of these human-robot collaborations as the targets of responsibility attribution in the event of harmful end results [41].

However, advocating for solely human-dependent responsibility has instigated arguments against the conditions creating-, and gravity of responsibility gaps [30, **?**], as well as arguments against the very existence of any kind of technologically-motivated responsibility issue [56]. Such arguments reject the call for scaling-down AI systems, including by those who say the responsibility gap can be bridged in the meanwhile, to instead argue that the foundations of moral responsibility are too fluid to leave any significant space missing [56]. Instead of the dissolution of responsibility or assigning it to the humans developing the technology, a milder approach uses a humans-in-the-loop method to identify three positions: "active learning, in which the system remains in control; interactive ML, in which there is a closer interaction between users and learning systems; and machine teaching, where human domain experts have control over the learning process" [39].

Nevertheless, others still have identified the issue of responsibility diffusion as worth taking seriously both in clinical decision support contexts [9], as well as in the use of intelligent tools and their constitution of an individual's environment [23].

Even more to the central point, newer accounts of the responsibility gap have made strides toward clarifying the specific domains for responsibility attribution. Some arguments track the gap as comprised of four interconnected issues of culpability, moral-, public- and active responsibility [50], while others focus on the various temporal aspects of backward-looking retribution as opposed to forward-looking responsibility attribution [16]. In questioning the assumption that only humans can be responsible agents, some emphasize the temporal aspect of the "many things" issue to link the agents of responsibility with the patients of their actions [13]. Using the notion of responsibility as answerability, such an account asserts that the patients receiving a decision may demand a explanation that answers for the costs incurred by the patient [13].

1.2.2 Explanation and Interpretation of Models

Two points of contention arise in response to these conclusions, namely: that the exercise of responsibility both requires and deserves AI experts who maintain a meaningful degree of control and oversight, and that increases in ML explainability may create an unwanted shifting of the loci of responsibility from agents to patients.

On the first contention, some have argued that for a notion of algorithmic transparency to be made useful to the public at large, we must employ the use of oversight bodies which enjoy full view of the data and machine learning development process in order to make a judgment of moral or legal responsibility [17]. On the second contention, three points have been established to argue how increasing explainability may shift responsibility to unwarranted patient parties; that AI systems providing post hoc explanations are occasionally viewed as blameworthy agents themselves, that variance among explainable algorithms can falsely imply patients have meaningful control, and that designers are truly the only group involved with any sense of meaningful control [32]. As such, there is a clear need for further exploration into the role that AI explainability has on the attribution of responsibility.

In comparison to the work on responsibility gaps, perhaps even more work has been done to offer various forms of explanation for machine learning decisions. To situate the ensuing discussion, is worth mentioning that interpretability has been argued to not be a monolithic concept, but instead constituted by several distinct ideas [33]. Driven by a desire for trustworthy, causality-inferring, domain-transferable, informative, and ethical algorithmic decision-making tools, the author argues for two properties of interpretable models: transparency and post-hoc interpretability. On the first property, transparency aims to identify how a model works, and is broken into three unordered levels of understanding – simulatability, decomposability, and algorithmic transparency – and this justifies reference to the central philosophical method of Levels of Abstraction

which will be used in this thesis. This provides solidity to the proposed method of analysis via a richer situation in the literature which, through further investigation, may benefit from clarification in these terms.

On the second property, post-hoc interpretability is a common taxonomic categorization for understanding how models operate after they have done so. Some authors believe that what little insight can be gained from post-hoc methods means that all attempts at interpretability must be qualified per model and can, in fact, mislead by offering plausible but incorrect explanations [33]. Others see explanations of the post-hoc variety as but one means of understanding how a ML model works via artificial interpretability, wherein such methods provide an understanding of how the system's inputs specifically combine to generate an output [48]. Relatedly, some models are themselves interpretable to a certain degree, via their hybrid neuro-symbolic architecture or inbuilt quasi-interpretable structures that mimic attention or map the salient features of the input. Nevertheless, some authors believe that no explanation of any such complicated models will replace the value and reliability of creating models that are intrinsically interpretable [49].

1.2.3 Understanding Via Models

Methods of understanding through artificial interpretability contrast methods of understanding without interpretability, which instead indicate only which properties of the input are deemed important for generating output [48]. An understanding of a model without interpretability can come through an analysis of the data, or instead using external support relations such as the strength of its relation to real-world evidence [55]. Though the coverage between understanding via artificial- and no interpretability is substantial, it is worth mentioning that some issues remain with regard to the nature of the explanations given, insofar as the pragmatic aspect of explanation is not fully furnished in these accounts [5]. Further, the precise requirements for understanding may not be the target of these explanations, but instead essentially distract with plausible placations. As such, the requirements for useful insight into these complex models will be borne out of the relationship between explanation and understanding [28].

Nevertheless, a division between artificial interpretability and understanding without interpretability is but one schema by which we might approach explanation in ML models. Among the viewpoints, some argue that because of a lack of consensus for normative criteria for dispelling algorithmic opaqueness, broad social acceptance of automated decision-making will only come domain-dependent strategies for explanation [15]. Alternatively, some taxonomies have been organized not around different domain strategies, but instead around the style of model at the heart of a decision; whether in machine learning or deep learning more specifically [4]. Others state that only a pragmatic account of understanding will offer progress and a useful starting point, and that approximation models are both the best means to understand an ML model and necessary for post-hoc interpretability [47]. There have also been attempts to characterize the opacity

facing the field [10], to make precise the nature of how machines themselves 'understand' and the epistemic consequences for our own understanding [3], or to develop formal frameworks for interpretable ML via idealized explanation games [59].

From this brief overview of some of the relevant literature, we can begin to articulate the central problem this thesis will attempt to clarify.

1.3 A Problem Emerges

Even from this brief description of the landscape surrounding the notion of understanding in AI, it appears that we lack a widely applicable framework relating a strong account of understanding to a means of normative comparison with which stakeholders may interpret models. The responsibility gap demonstrates the stakes of not being able to rigorously attribute normative qualities to the various aspects of complex AI function. Its multifaceted nature proves to be a difficult challenge for any set of explanatory techniques. To this end, we can identify some of the further questions which initially guided this research toward articulating a framework for understanding AI using LoAs.

1.3.1 Research Questions

This bridge between explainable AI and the plurality of notions for responsibility involved in deploying such systems is clearly not a trivial one to cross. One notable argument we will investigate further attempts to directly link the two sides of this issue [63]. However, the precise nature and degree of understanding of a complex model required to satisfy the explanatory requirements of various stakeholders remains unclear. One foreseeable issue involved in making general statements about the granularity of our understanding of a model and drawing normative conclusions lies in the fact that model architectures vary widely, so any attempt to describe "model understanding" must remain flexible across possible architectures.

With this in mind, the relevant research questions that have guided this project are as follows:

How might we effectively distinguish levels of transparency and their implications for comprehensiveness of understanding, such that we may use this understanding with regard to the various stakeholder commitments associated with its deployment? What account of understanding is most appropriate for linking it to normative criteria?

In the context of understanding a complex ML model, what criteria for understanding usefully establish normative constraints on our actions? If a model is to be built and used, to what extent must its inner functions be understood by stakeholders (including the developers, users, regulatory bodies, or otherwise) such that we can provide an explanation sufficient for responsibility tracing?

If a certain algorithm is to be deployed, how might LoAs help specify that a model's causal mechanisms are sufficiently interpretable? What contribution to our understanding is afforded by these Levels?

If it is true that the understanding we ordain from models is primarily limited by factors other than the LoA that has been specified (as is proposed by Sullivan's argument favoring link-uncertainty [55]), this seems to imply that much work in traditional forms of explainability will make little progress answering crucial questions about how such models work. More precisely, if unraveling a model's inner workings accomplishes little in comparison to grasping how a model's central concept maps to the target phenomenon in the world – as measured by the empirical support and linkage to the target – then understanding the model itself seems relatively unimportant.

So, we require some account of a model's function to provide an epistemological grounding to justify our decisions made based on their classifications. Toward this grounding, I argue that the underlying granularity of the Level of Abstraction by which one views a model has central influence on how it is understood and how it can be explained. Further, I aim to show that the structure of these LoAs is present in an important account of understanding, and that this can help produce novel insights related to the domain of AI. To begin, we must first consider this notable account of understanding and the ways in which opaque models can impede upon it.

2 Understanding and Explanation

Due to its centrality in the philosophy of science and other domains, the literature on understanding and explanation is vast. Thus, rather than attempt to overview the field, this description will be dedicated to a thin cross-section of selected writings on understanding and explanation for the purpose of explicating their role in the ML landscape. I begin by briefly discussing some key topics in the domain of scientific understanding including factivity, opacity, and the relationship between understanding and explanation. I then aim to articulate the role that explanations play in understanding complex processes being computed by advanced machine learning systems. The central proposition of this section is to establish the intermediary concept I have deemed The Relativity of Explanation. I conclude by mentioning a few important distinctions in the realm of understanding particularly as it pertains to the domain of machine learning.

2.1 A Close Relationship

A foundational account of the relationship between explanation and understanding comes is found in Khalifa (2017) [28]. Herein we are acquainted with the intimate relationship between understanding and explanation in Khalifa's Explanation-Knowledge-Science (EKS) model. With respect to empirical phenomena, this account details a model for explaining-why and thus is chiefly interested in explanatory understanding [28]. For the present purposes, an example of such an explanatory relationship would hold in the situation where "Abe understands why he was denied the loan", though of course such state-

ments need not include the term "why" verbatim. This account aims to establish a minimal threshold for understanding, as well as a comparative notion of explanatory depth.

Specifically, there are two related principles involved in providing an explanation which are necessary to provide a notion of comparison. These are known as the Nexus and Scientific Knowledge principles. These two principles aim to give a structured account of when one subject is more understanding of a statement than another subject. An explanatory nexus refers to the set of correct explanations for some proposition as well as the relations between those explanations ([28], pg. 6). As such, this web of correct and interlocked explanations can be grasped (by a subject) to different depths. Scientific knowledge on the other hand, refers to the accuracy of an explanation to real states of affairs in the world.

2.1.1 The Nexus Principle

The Nexus principle states that if subject S_1's grasp of the Explanatory Nexus is fully encapsulated by S_2's, we may conclude that S_2understands the topic better. Thus, we must determine i) what constitutes a correct explanation, ii) how the relations between them establish a meaningful form of understanding, and iii) what makes a subject's grasp more complete versus less ([28], pg. 6).

In the case of point (i), for an explanation of some statement to be correct, it must satisfy four straightforward criteria ([28], pg. 7). First, the statement must be (at least approximately) true. This should remain relatively uncontroversial, since there can be no reasonable explanation of why a rainbow is rectangular when it is in fact smoothly arched. Second, the explanans - the former statement doing the explaining - must make a difference in the explanandum - the statement to be explained ([28], pg. 7). Khalifa bases this difference-making notion on an acceptance of counterfactual dependence as a means of establishing that the explanandum is affected by the explanans[2]. More simply in Khalifa's terms, if event B counterfactually depends on A, then changes in event A cause changes in B, whereas a lack of change in A will not cause change in B [31], [61]. We can exemplify this notion by saying that the various lengths of electromagnetic radiation in light cause a difference in the scattering of various colors, and this difference creates the rainbow. Third, the explanans must fulfill the relevant ontological requirements. As such, this EKS account remains agnostic toward both realist and anti-realist modes of explanation. Therefore, we can proceed without reference to these further issues since aspects of this debate remain unresolved with regard to machine learning [25, 24]. Fourth, a proper explanation must satisfy local constraints, which allows for a plurality of explanatory methods to hold. As Khalifa admits, "the relevance of many explanatory features

[2]Note that counterfactual dependence is not the only way of establishing a difference, and that other means of difference making are also amenable to this account [28]. As such, the details of which specific notion of counterfactual dependence is used and how exactly it establishes causality should not significantly influence this account of understanding ([28], pg. 7).

depends on the specific explanandum, the standards of the discipline, and the interests of the inquirer" ([28], pg. 8). In sum, the first three criteria limit the scope of possible explanation whereas the fourth allows for some degree of flexibility.

Progressing to point (ii), to describe meaningful relations between these explanations, we can imagine a network mapping the causal dependency of some event worth explaining. If the nodes of such a network represent statements and the connectivity represents the statements explaining each other, of course it will be true that some explanatory routes through this network will be stronger and more relevant than others. However, rather than simply identifying which statements are stronger than others, a more comprehensive account of a strong explanation would come as a result of understanding the structure of such a network ([28], pg. 9). More precisely, a person would be in the most deeply-rooted epistemic position if they were able to expound not only on the strong routes through such an explanatory network, but also the interrelationships between these routes and how they undercut or support one another.

And finally for point (iii), the Nexus principle begets the strength of one's understanding by reference to the completeness of a subject's grasp. Completeness refers to the number of correct explanations as well as the interrelationships between them, their quality and importance, as well as the level of detail to which they are grasped by the subject doing the explaining. In the most simplified sense, a more complete explanation would typically encapsulate the explanatory nexus of one that is less complete. However, this is not meant to be a purely quantitative exercise. Providing a naive method of measurement to one's understanding as compared to another is not the ideal goal of this description. Rather, there will typically be "a stock of explanatory information that two or more inquirers both grasp and then some further bit of explanatory information that is unique to one," ([28], pg. 10).

So, we can imagine some empirical phenomena with a network of explanations which interact with one another to ground an explanation in truthful states of affairs in the world. We have already seen how grasping different aspects of this nexus relates to understanding, but now we will investigate the network's relationship to truthful states of affairs in the world.

2.1.2 The Scientific Knowledge Principle

The second comparative principle for determining relative degrees of understanding avoids reference to a phenomenological account of 'grasping'. Instead, it grounds itself in a notion of similarity to scientific knowledge. Therefore, the Scientific Knowledge Principle states that if S_1's grasp of a statement's Explanatory Nexus bears greater resemblance to scientific knowledge than S_2's, then S_1 understands that statement better than S_2 [28]. This leads us to wonder first, what exactly is meant by scientific knowledge and second, how exactly one's grasp bears resemblance to it [28].

First, to possess scientific knowledge requires that its holder have arrived at a belief that can only be the result of safe belief-forming processes that could

not have led to a falsehood [28]. Furthermore, the safety of this belief must rest upon a scientific explanatory evaluation. Such an evaluation is comprised of three main parts; initially, there must be consideration of the plausible modes of explanation, including modeling the explanatory nexus to uncover relationships between different explanatory factors. For our present purposes, plausibility refers to an explanation's fit with the relevant background theories as well as its simplicity[3]. Next, We need a means of comparing potential explanations to determine which are in conflict and which complement each other [28]. In cases where complementary explanations exist, we can inspect the Explanatory Nexus to uncover the differing dimensions along which various explanatory methods find their strength. To complete the safety-preserving evaluation, the prior comparisons should result in the formation of various attitudes about the relative strength and weakness of the explanations surveyed. As such, the comparisons between explanations should be sufficient for the investigating agents involved to assign them varying degrees of belief. This concludes our account of what constitutes scientific knowledge in this context; considering plausible explanatory options, comparing them to one another to determine complementarity, and altering the relevant agent's attitudes toward the explained phenomenon based on these explanations.

Second, resemblance is similar to the notion of completeness in the Nexus principle [28]. For an agent's grasp to resemble this account of scientific knowledge, it depends upon the number of potential explanations considered by the agent, as well as the number of comparisons made with methods that are acceptable by virtue of their scientific status ([28], pg. 13). The agent's beliefs about their explanatory nexus must be likewise safe and accurate as depicted by the previous discussion of belief safety and their fit with our discussed account of scientific knowledge. Furthermore, it is now worth noting the variety of possible cognitive states involving degrees of scientific knowledge. Where the standards for understanding are low, such that relatively little depth is required to justify a conclusion, Khalifa admits that merely approximately true beliefs may be satisfactory ([28], pg. 14). An increasingly accurate schema then results when we organize an agent's understanding from when it is based upon approximately true beliefs versus being based on scientific knowledge as described ([28], pg. 14).

2.1.3 The Explanation-Knowledge-Science Model and Degrees of Understanding

In addition to these two comparative principles, the Explanation-Knowledge-Science (EKS) model consists also of a notion of minimal understanding from which to begin meaningful comparison. Based upon the two criteria that have

[3]Khalifa adds that some potential explanations may of course be implausible, while some some plausible explanations may be incorrect. Though he provides no precise characterization of plausibility. This point is especially difficult in the realm of AI, where generating explanations means there may be no grounded means of verifying their truth. Our discussion of mechanistic interpretability in Section 4.2 will reconvene on this point.

already been discussed, minimal understanding of a why-explanation holds when a subject believes an statement explains some phenomenon and this statement is approximately correct ([28], pg. 14).

We can therefore summarize the EKS model by referencing its three main components and showing their arrangement. The first component is the Nexus principle which we can recall states that a subject may grasp the Explanatory Nexus of some proposition more completely than another subject [28]. Second, the Scientific Knowledge Principle states that one subject's grasp of such an Explanatory Nexus may bear greater resemblance to scientific knowledge than that of another subject [28]. Together, these two principles provide an account of what it means for one subject to have better understanding than another. Third, our aforementioned notion of minimal understanding states that, at the very least, a subject has minimal understanding of a "why" explanation for a proposition if and only if: that subject believes the explanation really does explain the proposition, and this explanation is at least approximately true.

Thus, there's a clear emergence of a spectrum of understanding readable within the Nexus Principle, Scientific Knowledge Principle and notion of minimal understanding. This overview can be briefly visualized to compare minimal understanding to scientific or more ideal understanding. On the one hand, minimal understanding would be a rather sparse, more weakly connected network of interexplanatory relationships. In contrast, the Explanatory Nexus of the relevant scientist would be richer, both in terms of the number and connectedness of its supporting explanations, but also in the subsequent Explanatory Nexuses of these supports, which further stabilize the scientist's understanding more than the minimal account.

There are two main reasons for including Khalifa's EKS account in this thesis. First, because it remains a relatively uncontroversial account of scientific understanding via explanation that is amenable to a LoA-based analysis - which will be made clearer in the following section - and second, that it reinforces the notion of explanation being agent-dependent.

Due to the ubiquity of understanding and explanation, the lack of controversy stems from the fact that this account is essentially "a more regimented descendent of the received view" ([28], pg. 16). This straightforwardly places this account of understanding in a central, rather acceptable position relative to the landscape.

Furthermore, this is an account of explanation that is dependent on the subject's abilities. It is worth noting that the present goal of explaining complicated systems need not require that increasing degrees of understanding are necessarily closer to an ideal account as opposed to a minimal one. Rather, we can at least establish that there are different forms of understanding pertinent to the various users who may interact with such a system. This is particularly important for the present purposes because of my advocacy for a concept I maintain as the Relativity of Explanation. Of course, there are accounts of understanding which are not relative to the subject. However, for the sake of explaining the decisions made by complex AI algorithms, it is sensible to make this distinction outright. This is due to the fact that various stakeholders interacting with

these algorithms bring various explanatory needs to bear upon them. As such , an account of understanding which was not agent relative seems to implicitly argue for a uniformity of explanation that would contradict the usefulness of the method to be proposed herein.

To create a useful example that will be referenced later in Section 3, we can imagine various different users interacting with a complex virtual assistant built upon an advanced large language model. A child may make a request to a virtual assistant to change the music, turn on the lights, or otherwise complete a relatively straightforward task. The child's parent may be a regular adult consumer more interested in shopping, planning trips, or automating their email system. One example of a more complete grasp of understanding could arise when the consumer parent is unable to book a ticket through this virtual assistant. Whereas the child might not understand why the process is failing, their parent may understand that they mistakenly provided the incorrect billing information. Thus, the parent could update this information and further, in choosing between providing one of two different credit cards, choose the one with a later expiry date so as to avoid the same problem in the future. In terms of an Explanatory Nexus, this would correspond to the parent grasping an explanation for the proposition (the billing mistake) and navigating their Nexus such that they could anticipate the proposition resurfacing. In comparison, the child may not be able to intuitively grasp the initial explanation, since they may not understand what a credit card is for. As a result, the Explanatory Nexus of the child is encompassed by that of the parent.

From this description, it may seem a naive quantification in reaching beyond another's capacity for understanding is the goal of this thesis. This is not the case however since we are less concerned with determining which of two individuals understand a why-explanation better than the other as much as we are concerned with fulfilling the explanatory needs of a given individual. Nevertheless, determining which explanations fulfill these requirements most strongly does involve a notion of comparison, and positing various imagined users proves helpful in demonstrating the fit of an explanation with the agent seeking it.

2.2 Opacity and The Relativity of Explanation

As we have demonstrated thus far, our central account of understanding is dependent upon a structured support system for some proposition being explained. This structured support system takes the form of an Explanatory Nexus, the accuracy of which is measured by the Scientific Knowledge Principle. Given this picture of how explanation supports understanding, we can now begin to develop a clearer image of the obstacles in the way of this support structure.

Up to this point, we have only discussed understanding when applied to abstract subjects insofar as we have not differentiated between different explanatory success criteria. As such, these vaguer caricatures do not accurately represent the possibility of a variety of stakeholders but instead some amalgam of capability for understanding. For the purposes of articulating a framework

for understanding when comparing to agents in the abstract, this will Suffice. However, we cannot make specific comments about the nature of understanding and explanation in AI systems without taking a more realistic view of the stakeholders involved in those explanations. Thus, we can now turn to an examination of the nature of opacity to provide further insight into how various stakeholders admit more specific normative criteria. More specifically, we can lay out the landscape of opacity in the field of AI to show how the understanding required for normativity interacts with those agents seeking such normative comparisons.

2.2.1 Forms of Opacity

One centrally important factor in providing useful explanations is that they be tuned to the agent seeking to understand. This forms the basis of the Relativity of Explanation, a concept I introduce to demonstrate an important connection between the users seeking explanations and the individual requirements therein. This Relativity of Explanation is supported by a central work on the concept of opacity in AI, a discussion of which shall form the basis of this section. We will now briefly examine an important taxonomy of AI opacity to set the groundwork we require.

Namely, Alessandro Facchini and Alberto Termine outline a three-part taxonomy of opacity in terms of **access**ing the internal workings of an AI model, its epistemic **link** to the systems being emulated in the world, and constructing a **semantic**ally coherent interpretation of the information within [19].

First, access opacity refers to the methods for dissecting the inner operations and behaviour of AI models to understand their function and structure [19]. This dissection is performed over the domains of the training dataset, training engine and learned model, which together are intended to exhaust the scope of an AI system's architecture. This form of opacity is most clearly identifiable when a human user is unable to locate and explain the functional role of the elements that are relevant for explaining and predicting various aspects of the system in question ([19], pg. 76). There are three identified causes of this epistemic access opacity [12], the interaction between two of which are most relevant to our depiction of the Relativity of Explanation. These two are the stakeholder's background knowledge and skills as well as the complexity of the system's structure, in terms of both its size and format[4].

Due to the natural limitations on human cognitive resources, our ability to properly conceive of a complex system's function diminishes as it increases in scale [19]. This applies not only to laymen with no domain expertise but also - to a lesser degree - to those involved in the construction of sufficiently complex systems. In particular, the training dataset may contain properties which make it difficult to gain epistemic access, but presumably even more dissection is required to understand the workings of the training engine and learned model.

[4]To make the same clarification as Facchini and Termine, size here refers to the number of the systems' elements on their mutual relations, while format refers to the type of elements included and their relations [34].

Specifically, there exists more innate complexity within the workings of the learning process than the block of data used to train the learning system [57]. This follows from the fact that the learning process depends not only on the data fed into it, but also the techniques for organizing the data with its corresponding outputs such that the model is useful.

Of course, this is not to say that the training data is without its own sources of opacity, the explanations for which would benefit from being user-relative. It is merely because to posit the Relativity of Explanation, the more impactful domains are those of the training engine and learned model. This is due to the fact that when a user encounters these complex artifacts, they cannot possibly seek to comprehend every aspect of its function simultaneously. As such, explanations for these two more complex artifacts of the system structure must adequately differentiate between the questions being asked of the system behaviour in order to provide relevant descriptions of the behaviour to be understood. It is this differentiation process which separates all possible explanations into those which cater to some users as opposed to others, depending on what they want to know about the processes within. Thus, the user's skills and background knowledge interface with the complexity of the system's structure to result in intersections which are the differently relevant questions. To account for this variability, the authors posit five different levels to provide these different descriptions "which may be suitable and relevant for some users but insufficient or inadequate for others" ([19], pg. 79). These five levels are Levels of Abstraction, which help to is the topic of the next section. This leveling framework is the culmination of the Relativity of Explanation, and it will be later argued that analyses based on such leveling frameworks are widespread in their applicability for deciphering the workings of complex computational systems. Thus, it forms a crucial aspect of this thesis.

Second in the taxonomy is link opacity. This refers to the degree of difficulty in correlating the mechanisms identified by the AI system with those it may mimic in the world [19]. If the AI system mirrors (and thereby explains) the mechanisms involved in some worldly process, we have learned something about the world system by reference to the mechanics of the AI system [55]. However, this may not be the case because of the heavy reliance on data driven ML approaches which do not operate using any hypothesis about the underlying patterns observed in the data, but which instead make broader probabilistic associations [55].

Whether an AI model is capable of furnishing mechanistic explanations likewise depends upon the mechanisms in the world conceivable by the relevant user. More precisely, the usefulness of a mechanistic explanation for a complex process in protein folding, for example, may fall short on any user that is not a domain expert. Thus, even the capacity to furnish mechanistic explanations beyond a sufficient depth required by the relevant user offers no guarantee that such a mechanistic explanation helps in providing understanding. Of course, a domain expert could reconstruct a more relevant explanation out of their understanding of the mechanisms and the information decipherable from the AI system, but this still requires an appropriate scoping of the explanation to determine how

the user receives it in their Explanatory Nexus and thus, understand[5].

The third category in the taxonomy is that of semantic opacity. Its two subdivisions can be briefly summarized for the sake of confirming that the Relativity of Explanation plays a role throughout the taxonomy. Before this summary however, we can note the two most relevant causes of semantic opacity. Specifically, these are when the model lacks a clear interpretive scheme which makes sense of the information it stores and inferences it makes, or when an available semantic for the learned model is incomprehensible due to the user's cognitive limitations, such as a lack of background knowledge or relevant epistemic skills ([19], pg. 85).

From these causes, the two resultant forms of semantic opacity refer to either the meaning of the information stored in the learned model (content opacity), or the inferences used to manipulate it (inferential opacity) [19]. In the former case, content opacity is most aptly demonstrated by the difference between rule based systems and neural networks. Here, logical sentences constituting the rules have their syntactic elements mapped to features that are relevant in this context by means of a standard Tarskian semantics ([19], pg. 86), whereas the same cannot be performed for the parameter weights in a neural network. In the exemplary case of medical diagnosis, a rule-based system would have its components such as predicates, variables, and connectives mapped to the medical evidence such as genetic mutations and disease history, but typically the parameters in a neural network are merely to reduce the overall error in prediction, not that any of them are individually interpretable [19]. In the latter case, inferential opacity refers to the inability to make sense of the reasoning paths followed in one of these decision makers ([19], pg. 87). Like with the values of neural network parameters, the inferences may only have an instrumental value insofar as they are accurate in prediction.

Once more, note the persistence of the user's limitations in causing semantic opacity in both of its forms and - more importantly - how there is an implied degree of detail assumed to be relevant in determining which aspects of the informational structure helpfully receive a semantics. For example, the transistors and instances of memory access within a computer involved in producing a simple neural network classification are typically not the items deemed worthy of requiring a semantics. Instead, the neurons themselves are assumed to be the items deserving of a semantic interpretation. In other words, the components within the functioning AI system are not all equally demanded to have a semantics applied to them, and thus there is an implied constraint on the degree of detail presupposed to have an influence on understanding the information in the model. As such, it appears as though there is an implied choice of LoA

[5]More can be said about the relation between this type of explanatory scoping and link opacity. However, for the purpose of this section in asserting the Relativity of Explanation, it may suffice to briefly mention that the notion of matching mechanisms within an AI model with those in the world being discovered is similarly dependent upon an initial set of scoping decisions. A fuller justification of this point may be deserved, however it must wait until after a deeper discussion of Levels of Abstraction (read scoping decisions for now) in the next section in order for it to be coherent.

being made which is simply not deemed worthy of explicating. This begins to motivate the necessity for including LoA considerations beyond the scope of only access opacity.

2.2.2 Relations Between Forms of Opacity

From the stakeholders perspective, this taxonomy supports the notion that various stakeholders bring diverse explanatory requirements to bear on the system to be explained. As such, the context in which these AI systems are deployed, which incorporates the users objectives, background knowledge, and cognitive abilities, directly influences the manner in which the system is perceived as opaque. The Relativity of Explanation therefore has a fundamental role in shaping the form of this taxonomy on opacity, such that it is characterized as a context-dependent and plural concept whose usefulness depends on its ability to cater to these diverse explanatory needs.

So aside from our overview of this tripartite taxonomy of opacity, the notable subdivisions within access opacity in particular lead into our next topic. To begin transitioning toward our discussion of Levels of Abstraction, it is important to recall the pervasive importance of appropriately scoping the potential explanations for the stakeholder. This has laid the groundwork for the Relativity of Explanation as I have posited it. Toward the major goal of this thesis, the Relativity of Explanation represents a bridge between the inscrutable data-driven AI models and the set of established standards for explanation and understanding by focusing on stakeholder explanatory requirements.

To transition from this notion to the next section on Levels of Abstraction, we can recall how the alignment of a user's explanatory criteria with the complexity of the system to be explained results in a limited set of possibly useful descriptions. These sets of useful descriptions can be organized into a level-based framework in a way where the levels represent cross-sections of a system can be viewed independently of one another. These levels reflect a given user's explanatory requirements and their arbitrarily definable set of possible questions of the system's function. Further, they can provide unique information depending on the method of their construction.

In the case of this taxonomy and its handling of the training engine and learned model artifacts (under the umbrella of access opacity), the authors repurpose five levels for organizing the functionality of the hardware and software used to create these artifacts. The higher levels are the more abstract and they are concerned with the complex behaviour of the software in the system, while the lower level details eventually terminate in hardware processes computing everything underlying those behaviours. Each level passing from software down to hardware represents a different lens through which to attempt viewing further through the opacity of the system. Of course, the organization of these five lenses on opacity as levels is no coincidence, since the functioning described at one level is closely tied to the functioning described in adjacent levels.

As has been the main argument of this section, these levels are intimately tied with specific user explanatory success criteria. In particular, having a detailed

understanding of advanced software behaviors serves a different set of potential users than an understanding of every physical piece computing those behaviors.

One critical observation included by the authors involves the dependencies between these forms of opacity. In particular, it's worth noting that the only explicit reference to LoAs in this taxonomy happens within the training engine and learned model subdivisions under access opacity. However, the authors go on to explain that access opacity in a learned model can "cause link opacity whenever the user's epistemic access to the LoA providing the information that is relevant for the understanding of the target phenomenon is limited" ([19], pg. 87). Further, we can recall that a user's ability to provide a semantic interpretation of the relevant aspects of a learned model is foundationally involved in semantic opacity, but "this ability may be compromised by a limited epistemic access to the concerned LoAs and therefore cause semantic opacity" ([19], pg. 87). As a result, it seems that there is reason to reevaluate the role of LoAs as an analytical method beyond the five part hierarchy identified within access opacity. It is this intuition which motivates the broader usage of an LoA styled analysis for understanding not only the systems in question themselves, but the relations between LoA choice and other forms of opacity.

Outfitting a user's Explanatory Nexus with the relevant information requires an appreciation of the nature of that user as a stakeholder in the system being explained. As such, interlocking the user-specific dependence of the Relativity of Explanation within the form of the Explanatory Nexus begins to ground the use of these Levels of Abstraction as explanatory tools for understanding. To fully understand this process however, we must now develop our account of these LoAs further.

3 Levels of Abstraction

This section will very briefly describe the usefulness of a Level of Abstraction (LoA) within the context of epistemological analysis. After situating how an LoA relates to the systems it analyzes and the models it produces, I detail how an LoA is constructed and used (Section 2.1), how they differ from similar approaches to understanding (Section 2.2), and how a LoA-based analysis of ML systems can enrich the normative framework landscape beyond the current state (Section 2.3).

Broadly speaking, an LoA is a framework for understanding some system by essentially taking a specific view of it. More specifically, this framework is a form of epistemological levelism which attempts to explain phenomena in the world with reference to various, possibly intersecting descriptions [20]. These descriptions are often layered upon one another such that the nested LoAs provide increasing degrees of detail. They are epistemological insofar as they aim to study reality at various levels for the sake of breaking complex phenomena down into understandable components. As such, this framework refrains from making ontological claims about objects described at different levels, or from articulating methodological interdependence among layered theories. Instead, LoAs are

a means of interpreting some variables within a system, whether conceptual or empirical. The choice of variables within the LoA determines the nature of the models which may be generated from it, which themselves are intended to identify some structural regularities within the system under consideration [20].

By choosing some aspects of a system to be variables, the designer of a LoA is considering only these aspects of the system and all else is abstracted out. This means that those entities which are epistemically relevant to the system processes under consideration are chosen as the foundation for the models made thereafter. As a result, the ontological commitments of the LoA are borne out after the variables are decided[6]. This simplifies the process of analysis to allow for more streamlined understanding of which questions can be meaningfully asked of and answered by a given model [20]. Through the process of analyzing a system at some LoA and generating a model within the confines of the variables chosen, the boundaries are set as to what information is within the scope of the LoA.

To begin understanding the LoA framework in the broader context of the goals of this thesis, we can consider that one of the simplest descriptions of a LoA is as a specific perspective on a system under consideration [20]. Under this description, we might imagine various stakeholders of a complex system and their different views or judgments on the subject matter, such as whether to engage with it. For example, the different users of an advanced virtual assistant would have different criteria guiding their understanding of it. A child receiving their first phone may only know that the assistant has certain features which make rich interactions possible and may not care about the inner mechanisms driving these interactions. However, the assistant is still required to perform many functions that may go unappreciated by the child. The child's judgment would correspond to a highly abstract LoA insofar as their criteria for choosing among assistants may be a rather limited set. As such, variables relating to ease of use and basic functionality might suffice.

Still at an abstract LoA but different from that of the child, a more common consumer might value certain variables such as the ability to make online purchases, plan a travel itinerary, or automate the sending of emails. There may be some personal enjoyment in understanding the inner language model producing the assistant, but certainly it is the factors involving its practical household status that determine its worth to most consumers.

To distinguish an abstract LoA from a more granular view, consider the relevant criteria for a software engineer or researcher. The engineer would likely be interested in the details of the assistant's interoperability with software through the relevant APIs, the processes by which the assistant breaks down and re-

[6]Floridi distinguishes between a committing and committed component within a system-level-model-structure (SLMS) scheme to show this process of a theory explicating its ontological commitment. Briefly, this SLMS scheme shows a cycle of relations where some System is analyzed at a LoA, which generates a Model to identify a Structure attributed to that system ([20], pg. 316). Thus, when a theory (encompassing the level-model-structure components) accepts a LoA, it commits itself to the existence of certain types of objects, which are the types constituting the LoA observables. Since this metaphysical picture is beyond the present scope, the entirety of this process can is articulated in [20], pages 315-316.

sponds to a prompt, or its manner of navigating the internet to answer complex queries. These lower LoAs are constituted by variables which give more detailed views of the internal functioning of the assistant, rather than only its outward-facing capabilities.

From these simple cases of distinctions among different types of users, we can already begin to see the difference between a perspective which values a virtual assistant as something which makes shopping more convenient versus as a tool for accomplishing specific tasks. This allows us to begin building an intuition around how the different variables chosen by users shape each of their views as to the resultant value of the same object.

With this brief sketch in mind, we can now delve into a more detailed look at the components constituting a LoA beyond their rough contours, as well as how they relate to one another and other similar work.

3.1 Definitions and The Method

As mentioned, the first component of a Level of Abstraction is that of a typed variable. Such entities will be familiar to all who have studied computer science, insofar as a variable is a placeholder for a referent which may be unknown or changeable [20]. The following quotations are taken from the original account in Floridi (2008).

Definition. "A *typed variable* is a uniquely-named conceptual entity (the variable) and a set, called its type, consisting of all the values that the entity may take. Two typed variables are regarded as equal if and only if their variables have the same name and their types are equal as sets" (pg. 305).

For the purposes of notation, $x{:}X$ designates a variable x of type X. This means that any variable has a set of predefined types, which constitute an important decision about how that variable may change with reference to changes in the observed system. In fact, once a variable has been stated with a typical place holding symbol, and given a restricted set of possible values, it can be interpreted such that it becomes an observable.

Definition. "An *observable* is an interpreted typed variable, that is, a typed variable together with a statement of what feature of the system under consideration it represents. Two observables are regarded as equal if and only if their typed variables are equal, they model the same feature and, in that context, one takes a given value if and only if the other does" (pg. 306).

Further, observables are *discrete* if their variables have finitely many values, and are *analogue* otherwise [20].

Although in some contexts it is appropriate to leave implicit the relationship between the observed feature of the world system and its corresponding feature in the model, it is important to explicate this relationship for the sake of clarifying the process of inference. Generally speaking, a system can refer to any object in the world worth analyzing, including conceptual systems whose

domain is discourse, analysis, or other purely semantic grounds. However, for the purposes of this thesis, a system will typically refer to a machine learning algorithm instantiated on some computing device and will be disambiguated from the world system being predicted otherwise [7].

To clarify the relationship between observables and typed variables, we can imagine measuring the height of a person [20]. If the variable for a person's height is assigned the letter h, and the set of possible types given is the rational numbers such that $\{0 < h < 3\}$, this provides an example of a typed variable. If we add the statement that h represents that person's height in meters and interpret this value as such, this makes it an observable. Note that the interpretation must be consistent, since interpreting the a similar variable in feet and inches may appear to equate the typed variables despite this being an error.

With a clarified notion of an observable in mind, we can now construct a more precise definition of our central LoA concept, which is essentially a collection of observables.

Definition. "A *level of abstraction* (LoA) is a finite but non-empty set of observables. No order is assigned to the observables, which are expected to be the building blocks in a theory characterised by their very definition. A LoA is called *discrete* (respectively *analogue*) if and only if all its observables are *discrete* (respectively *analogue*); otherwise it is called *hybrid*" (pg. 309).

We can now briefly return to the previous examples of various computer users. The child LoA would presumably consist of observables capturing ease of access such as "display size and resolution", whereas the engineer would likely focus on observables for "storage capacity" or "processing capabilities" as described previously.

Thus, observables alone describe the aspects of the system worth consideration as independent entities. However, the system being considered likely cannot take an arbitrary assortment of possible variable values and remain realistic. So in addition to a collection of observables, LoAs need restrictions on all possible combinations of values to understand the relationships that are allowed to hold among the observables within a given LoA [20]. For a straightforward example, we can take the case of a person's height where the variable type is the rational numbers. Any value inside the given range $\{0 < h < 3\}$ would be considered a possible system behaviour.

More sophisticated than a restriction on a single value, consider the properties of a chess game. Typically during a match, players simply record their moves in a standard notation. Potential observables could be the time taken to perform a move, the apparent stress level of the players, their exercise habits in

[7] Nevertheless, there will be cases where the distinction between the world system being modeled and the algorithm designed to recognize patterns within it remains unclear. Especially in cases where machine learning models are identified as such, the explanatory work may describe the machine learning model as the system (object) in question. Therefore, the use of the term "system" will be used here to refer to the object being studied. This will typically be a machine learning system, not the system in the world that the machine learning algorithm is attempting to predict or otherwise emulate.

training, or the location of the match. This information may not appear directly relevant for understanding how the pieces move across the board, yet these features may inform why a particular blunder was made or good move overlooked. The annotated moves in particular are only informative with the relevant background understanding of the initial and following possible states of the board. To make this understanding explicit, we can disassemble the constructed game into various constituent pieces[8].

Imagine viewing along the surface of the chessboard from two perpendicular angles. The first LoA would only view the files of the board stretching from player to player without consideration of the ranks running from side to side nor the differently colored squares within [20]. The second LoA would only consider the ranks and would not distinguish the files. In the first LoA, each file's observable would thus be the set of pieces placed somewhere upon the eight squares in it, and a move would be constituted by the change in file that results. Observing each piece as it moved would show a knight moving one or two files to either side, a king could only move one file side to side while a bishop could move any number of squares. A rook (or indeed any other piece) that moves along the file would appear stationary in this view, but would be indistinguishable from a bishop if it likewise moved between files. Similarly, a pawn could only change files when it captured another piece, otherwise it would traverse down its file and appear to never move. Pawn movements in the second, rank-based LoA would always show a change of position since pawns must move forward.

Where neither view of file- nor rank-based chess would individually provide much useful information, "the two disjoint observations together ... reveal the underlying game," ([20], pg. 308). This is due to the fact that the game has been disassembled into two dimensionally-impoverished LoAs whose true value is realized once they are recombined to recover the original two-dimensional game. The standard notion again becomes meaningful, since the rank and file coordinates of each piece which were projected down into each LoA in the deconstruction process.

However, some information remains lost if we attempt to understand the original game from only this pair of LoAs. As mentioned with the example of a person's height, we still require some restrictions on the possible observable values that permit us to make sense of the overall system. Otherwise, an unconstrained set of observables taking arbitrary values would not usefully correspond to the system in question. To this end, we can define a system's behaviour as precisely those combinations of observable values that correspond to real states of affairs.

Definition. "The *behaviour* of a system, at a given LoA, is defined to consist of a predicate whose free variables are observables at that LoA. The substitutions of values for observables that make the predicate true are called the *system*

[8]Of course, there are many possible methods of performing this dissection, including whether to only include features of the chessboard itself or these other contextual features. The goal is to create the most useful simplifications which remain maximally informative to the intentions sought by the LoA designer.

*behaviour*s. A *moderated* LoA is defined to consist of a LoA together with a behaviour at that LoA" (pg. 310).

Thus, a moderated LoA couples a specific view of a system to its allowable dynamics. This tracks the representation relationships between observables and their instantiated aspects within the system in question. This is particularly useful in discrete systems as opposed to analogue, since the behaviours of analogue systems in science are typically described by differential equations. To appreciate the relevant difference between discrete and analogue systems, it is worth identifying their respective domains of application.

Most importantly, the continuity that holds within analogue systems means that small changes in observables create small corresponding changes in the overall system behaviour ([20], pg. 310). The quantities of these changes can be described exactly and solved, providing the sought behaviour in a way that cannot be achieved in discrete systems. Instead, where discrete systems have observables whose small variation may cause arbitrary resultant changes in the overall system behaviour, the method of LoAs provides a way of tracing the relationships that hold across these apparently arbitrary changes ([20], pg. 310). This allows the system to be comprehended via simple approximations; in this case, the predicate's free variables constitute an exact description of the system behaviours. Then, varying the LoA permits increasingly detailed accounts of the system while retaining the correspondence from observable changes to system behaviour changes.

We can now arrive at the final remaining definition required to understand the method overall. In contrast to the scope of a single model as formalized by a LoA, a Gradient of Abstractions describes the process for facilitating discussion over a range of possible LoAs [20].

Definition. "A *gradient of abstractions*, GoA, is defined to consist of a finite set $\{L_i | 0 \leq i < n\}$ of moderated LoAs L_i, a family of relations $R_{i,j} \subseteq L_i \times L_j$, for $0 \leq i \neq j < n$ relating the observables of each pair L_i and L_j of distinct LoAs in such a way that:

1. the relationships are inverse: for i = j, $R_{i,j}$ is the reverse of $R_{j,i}$
2. the behaviour p_j at L_j is at least as strong as the translated behaviour

$$P_{R_{i,j}}(p_i)p_j \implies P_{R_{i,j}}(p_i)$$

and for each interpreted type *x:X* and *y:Y* in L_i and L_j, respectively, such that *(x:X, y:Y)* is in $R_{i,j}$, a relation $R_{x,y} \subset X \times Y$," ([20], pg. 312). The same equality and discrete/analogue restrictions apply as with all previous definitions.

Condition (1) ensures consistency is maintained between successive LoAs as increasing detail is sought. In particular, if new observables are added to a more detailed, lower LoA which thereby extend another higher LoA, the same

constraints operating on the higher LoA apply to the lower. Even if newly introduced observables lie outside the scope of the more abstract LoA, the constraints applied to these observables are still true in the more abstract LoA.

The goal of the method is therefore to adjust the LoAs such that they become more comprehensive in their detailed expression of system behaviours [20]. The variability within LoA construction is a core feature of the method which plays a key role in distinguishing the herein proposed LoA method from other normative frameworks.

3.1.1 Disjoint and Nested LoAs

One useful distinction between the possible forms of a GoA is between their having disjoint or nested LoAs. A disjoint GoA is one with constituent detailed LoAs containing complementary information to one another which combine to give a fuller account of the more abstract LoAs. In contrast, nested LoAs provide further refinements on the details of a system when they are each successively within the scope of the previous. More specifically, we can turn to the defined difference between these two options:

Definition. "A GoA is called *disjoint* if and only if the L_i are pairwise disjoint (i.e., taken two at a time, they have no observable in common) and the relations are all empty. It is called *nested* if and only if the only nonempty relations are those between L_i and L_{i+1}, for each $0 \leq i < n-1$, and moreover the reverse of each $R_{i,i+1}$ is a surjective function from the observables of L_{i+1} to those of L_i" ([20], pg. 312).

Most commonly useful to the domain of computer science are the nested layers. Their hierarchical structure offers a successively fine-grained representation of the computation being performed. By using a nested GoA defined to capture the behaviour of a system, we can provide an example of the nature of the relations between related LoAs. In what might be the simplest case, we can recall the example of a person's height or we can measure their physical attributes as a proxy for their potential aptitude for some sport. Specifically, LoA L_0could contain observables for a person's size - designated by variable s, typed with {small, medium, or large} shirt sizes - and the dynamism in their movement - designated by variable d, typed as a real number representing the time in seconds it takes for them to move through a standardized set of markers on the floor. A more precise LoA L_1might then contain two more useful variables for both size and dynamism, which receive reach and height, and ground speed and jump touch respectively. Measuring one's reach and height as representative of their size could be done with variables r and h, each typed with bounded[9] real numbers representing centimeters to provide the behaviour of L_1as the following predicate with free variables r and h. This system behaviour

[9]In this case, the upper boundary could simply be the value of the height of the tallest person ever recorded [58]. For the purposes of this example, this provides an allowable range of physical measurements that could indicate a person's aptitude for success in some sport.

could be interpreted as some threshold of practicality for being considered for a competitive team.

$$\{r_{min} < r < r_{max}\} \wedge \{h_{min} < h < h_{max}\}$$

In summary, a LoA is a set of variables - each with a specified type - which are the interpreted aspects of the system under observation[10]. Specifying a LoA helps to clarify which questions are meaningfully oriented toward the system at hand, as well as to disambiguate and clarify fallacies and category mistakes. This is intended to establish the rules by which the subsequent analysis is governed; as Floridi states "the choice of a LoA predetermines the type and quantity of data that can be considered and hence the information that can be contained in the model"[20]. As will be shown in Section 3.4, this variation in the information that can be contained within a model reflects how LoA-change contributes to an understanding of the target system.

3.2 Similar Approaches Toward Normativity

Using LoAs to clarify conceptual commitments in a model is a fundamentally useful epistemological method, but it is not altogether new. There are multiple notable methods that attempt to clarify how one may take differing views of a single system and extract meaningfully different information from each. As such, this section will discuss various methods of interpreting the behaviour of complex systems via LoA-like analyses, each with the purpose of being internally comparable with reference to various normative commitments. Section 3.3 will discuss precisely what is meant by these frameworks having normative targets in more detail, and Section 3.4 will combine the full account of how a LoA-based approach offers key insights into understanding the complexities of AI systems and their associated normative commitments.

One of the most famous attempts at a framework for understanding complex systems through layers of analysis is provided by the neuroscientist David Marr [35]. Here, Marr posits three complementary views on how a complex system performs the task it is designed to accomplish: the computational, algorithmic, and physical levels. These give increasingly fine-grained descriptions of the component pieces within the system interacting to perform its overall task.

A further attempt at understanding the field of AI is sketched by Serb and Prodromakis (2019), who add two layers atop Marr's system in an attempt to provide a more exhaustive coverage of possible descriptions [52]. Additionally, they aim to identify pathways for potential innovation within certain safety and performance guarantees as described by the framework's levels. What follows will establish the fundaments of Marr's framework and briefly expand upon a cross-section of the dialogue surrounding it's relevance to the digital domain. In

[10]For another account of structured views of complex systems that approaches a similar goal via algebraic topology to explain spatial and structural changes over time within a similar hierarchical system, consider Atkin, 1982 [2], as well as Beaumont and Gatrell's discussion of his method in [7].

particular, we will overview some attempts to make this framework amenable to the tasks associated with understanding advanced AI. After situating Marr's framework in some of its contemporary discourse, we may proceed to a comparison of the various innate tradeoffs within these different leveling systems.

3.2.1 Marr and Machine Learning

Perhaps most famously, the vision neuroscientist David Marr aimed to dissect complex systems using three complementary levels of analysis [35]. Despite originally being construed as a means of understanding vision processing in the brain, these levels have become foundational pillars in complex system analysis more broadly. Thus, it is helpful to begin with this oldest framework to fully understand the later attempts at making Marr's levels amenable to ML systems. In this way, we can begin to understand the proposed value of analytical methods that depend upon such level-based frameworks and the benefits they provide. These benefits will be briefly overviewed after we are introduced to one possible expansion of Marr's framework from Serb and Prodromakis (2019) [52], and a further discussion of the benefits of leveling-systems will come in Section 3.4. Generally speaking in the domain of AI, the clarifications offered by this style of analysis allow for simplified comparisons between interpretability taxonomies, identification of various forms of ethical risk, as well as more accurate potential behaviour prediction.

Due to its centrality in fields such as the neuro- and cognitive-sciences, Marr's threefold framework is renowned for its popularity. We can sketch its layers of analysis to serve as a proper starting point for comparison with other similar theorizing. Specifically, this framework consists of three distinct levels, all of which describe a complex system with increasing implementational detail. From the top down, Marr's levels are the computational, the algorithmic, and the physical [35]. These levels correspond to three descriptions of a system which elucidate particular features depending on the view assumed.

Most abstractly, the computational level addresses the broader questions related to overall task decomposition and processing. The main components of the task and their manner of serving the overall goal of the designed artifact are the foci of this level. This level seeks to define the main purpose of the process, and what kind of information it requires in order to be accomplished [35]. Stated in terms of the original vision example, the computational problem of recognizing objects would be how a visual system extracts information from the sensory input to make categorizations of objects and their relationships in space [35].

Next, the algorithmic level aims to specify the steps which constitute the algorithms used to achieve the goals set out at the computational level [35]. As such, this level articulates the manner in which the incoming data is processed into its desired output state. This requires a description of the representations and various transformations of the information involved in all the intermediary steps. To use the same example as earlier, we can recognize objects by first addressing questions of how edges might be detected in the visual field, how

contrast and contour allow structure to be determined, or how we can perceive depth by differentiating between the information given by the two points of input in our eyes [35].

Finally, the most concrete level is that of the implementation of the previous level within the physical hardware [35]. The physical level describes how the algorithms identified are made manifest within the biological, digital, electronic, and other mechanisms underpinning the actual completion of the algorithms. Following suit, this level describes the specific circuits involved in transmitting the signals that are processed in every intermediate step of the algorithms. Again with our example of recognizing objects, the physical level would describe which neural cells and pathways are involved in the various aspects of transmitting the visual information from the eye's first input, to the recognition that an object is in the visual field [35].

Much has been written about Marr's framework, but it is worth summarizing a few key points before we discuss the expansions upon his theory which aim to apply this framework to the domain of machine learning. One straightforward observation is that there exist certain dependencies between the levels as they are described. There is a conceptual flow as the more abstract levels become more concrete. More specifically, the goals set out in the computational level carry through to determine the structure of the algorithms which furthermore bound their physical implementation to some configuration of components executing them. In this sense, there is a top-down flow of information which determines the criteria for evidence of success from the computational to the physical level. Likewise in reverse, the performance constraints of physical hardware can influence the usefulness of various algorithms. The availability of sufficiently complex algorithms then constrains the range of possible goals that can be accomplished. Thus, there is a bottom-up flow of constraints imposed by logistical realities of solving problems.

Moreover, there can be cases where the boundaries between levels blur, since the system under investigation may have many complex components each with their own definable computational-level goals. For essentially this reason, Marr advocates for an iterative application of this system at varying spatial scales [35]. This is intended to enable the framework's usefulness to be renewed regardless of whether the system in question is the entirety of visual processing, a certain brain region firing in sequence, or a single neuron. In each case, we can define its computational goal, its algorithmic processes, as well as their physical implementation[11].

Recently, Marr's levels of analysis have been argued to offer uniquely distinct contributions when analyzing not only biological vision systems, but especially information-processing mechanisms more broadly [8]. In fact, the necessity of each level in the framework is derived from those unique contributions to the

[11]This blurring of the boundary between levels is particularly evident with the advent of wetware biological computing which breaks down the typical hardware-software distinction. For a more detailed account of the transference of abstractly defined specifications into this genetic circuitry "between" the hardware and software, while maintaining the distinct usefulness of these different levels, see (Oliveira and Densmore, 2022) [45].

analysis. More specifically, the computational level helps to conceptualize why a mechanism does what it does, and places this overall notion of its function within its broader environmental context [8]. This helps build an understanding of the external tasks and constraints into an answer of its contextualized purpose. Internally however, the algorithmic level inspects the encoded representations within the mechanism to clarify how information is being processed. Bechtel and Shagrir argue that this "how" of computation defines the manner in which these patterns of organization enable the mechanism to produce its particular phenomenon [8]. Finally, the implementational level completes this top-down picture of necessary levels since it provides a comprehensive view of how the mechanism is performed in reality, along with the physical details constraining its operation [8]. As mentioned previously, these levels interact in their contributions such that the computational level defines the overarching goals to be accomplished, the algorithmic level explains how these goals are achieved via internal processes, and the implementational level produces the result in reality. So with regard to the methodologies identified by each level of analysis, we see how each perspective offers a different understanding of both the internal workings of the functions and the overall trajectory of the system at hand.

Moreover, Marr's levels have been used to analyze various types of ML systems to make precise the nature of the algorithms involved in reinforcement learning (RL) and beyond [40], [22]. Broadly, Hamrick and Mohamed (2020) take elementary steps toward directly applying Marr's framework to various methods in ML. First, they identify one example of how deep Q-networks (a type of RL algorithm) could face what appears to be one critique - namely, that of the algorithms failing to really know what objects are - but which breaks into two separate critiques when made at two different levels of analysis. More precisely, the computational goal is to maximize scalar reward, but perhaps this should be formulated with direct reference to understanding objects and spatial relations. Likewise the implementational critique could be directed more at the manner of distributed representation used by these Q-networks, since there is a potentially misguided translation from the discrete composition of real objects to these distributed forms [22].

To demonstrate the scope of possible AI domains to which Marr's levels apply, the authors also use the framework to articulate the relationship between symbolic and distributed reasoning systems. For something such as the Traveling Salesman Problem, solutions are traditionally implemented using symbolic programming methods although, algorithmically, the TSP can be converted to other NP-complete problems which can themselves be solved using a variety of algorithms, including heuristic search [22]. The programming implementations of these heuristics need not be symbolic however, and as such they could be implemented with non-symbolic deep learning components. As a result of this analysis, we see that ML systems may enact computations using both symbolic and distributed representations at differing levels of analysis [22].

Furthermore, the application of Marr's levels can emphasize crucial design considerations such as choosing how to represent discretized time brackets, to ensure that the logic of the algorithmic level correctly achieves the computational-

level goal, as well as to track the subtle goal changes made between levels [22]. Likewise with bias in the training data, where these levels might help distinguish biases as the "result of clinical constraints at the computational level, or whether it is an artefact of algorithmic-level choices"[22]. Notably, the authors also mention the flexibility of Marr's leveling system insofar as it can focus the analysis on various aspects of larger systems. However, they are careful to add that Marr's framework is unable to "capture the full set of abstractions ranging from the computational-level goal to the physical implementation" [22]. And more broadly, they acknowledge that Marr's levels lack any consideration of the socially-situated role of computing systems. As such, we seek a framework that maintains the insightful cross-cutting nature of Marr's levels, but which is more flexible and responsive to contextual concerns.

More elaborately than the brief RL example previously, Niv and Langdon (2016) argue that the entire field of RL has had enormous success due to its straddling of all three of Marr's levels [40]. However, they assert that especially in the domain of RL, open problems remain when applying Marr's levels of analysis and that the solutions to these problems require input from various levels for their resolution.

For example, the computational level goal of these decision-making systems is to maximize their future reward and minimize their punishments, but this might not be a true representation of agent goals from the perspective of evolution [40]. As such, we may ask whether this formulation is most apt for the tasks we seek to achieve, or whether more nuanced forms of fundamental rewards should also be considered such as curiosity and information-seeking [40]. Similarly, task representation is typically learned via experience, but the methods by which this is done are not clear. Where animals have been observed to infer the causal structure of a task based on their observations, the question remains as to what mechanisms are involved and how they relate to the agent's memory and attention.

Algorithmically, the representation of tasks as temporally distributed has been used to help navigate continuous learning, but this begs the question of how these representations adapt to different learning environments while accounting for the passage of time in complicated tasks [40]. Likewise, internal action and state representations can determine the efficiency of many RL algorithms in hierarchical-task scenarios, especially insofar as the decomposition of these task hierarchies lack proper criteria for which modes of decomposition support learning the most [40].

At the implementational level, successful mappings from RL functions to neurobiological substrates leave unanswered the question of how certain brain neurons compute reward prediction errors as analogized in the RL algorithms. Finally, a combination of model-based and model-free architectures may coexist in the brain, but this leaves open the question of how such systems interact with one another or the overall system behaviour [40]. At the very least, these findings motivate the use of LoA-styled analysis for understanding a variety of advanced ML architectures, especially to diagnose misplaced assumptions and make clear which analogies are relied upon too heavily. Thus, we require a

framework that maintains the usefulness of Marr's approach and which may recognize the broader context of computing systems and their interactions with users and other computing systems.

In accordance with the recommendations made by both Hamrick and Mohamed as well as Niv and Langdon, in the next section, I will discuss the need for and benefits associated with a more fine-grained approach than that offered by Marr's framework. Rather than only seeking to find inspiration across the levels, the LoA method will demonstrate a more flexible and modular framework than that presented in Marr's three levels, while maintaining its usefulness.

3.2.2 A Five-Part View of AI Systems

First, the uppermost layer in the Serb-Prodromakis system is that of agency. At this level, we can interpret the system as exercising the ability to make goals and decisions, learn to improve its behaviours toward those goals, and acquire declarative knowledge to act more effectively in its environment [52]. This layer is perhaps most characteristic of human intelligence and it plays a key role in understanding our interactions with such systems, since it allows us to formulate an idea of a machine whose actions are reminiscent of our own. This layer is therefore also crucial in bounding an AI system's behaviour within some set of ethical norms [52]. As a result, research in cognitive psychology, the ethics of human-computer interactions, and multi-agent systems all play a role in enhancing the agential capabilities, to name a few.

Next, the semantic layer deals with the manipulation of symbols, as well as more fine-grained reasoning and planning. Although this layer also deals with aspects of reasoning, its particular operations consist of those such as variable assignment, inference, and the use of memory - both for storage and recall [52]. Often, the messages communicated between component pieces at this layer occur using semantic objects which are typically represented as high-dimensional vectors [52]. Typically, learning at this level consists of the transformation of these semantic object vectors to form new objects from salient combinations of the old. In contrast to the lower layers which may rely on deep neural networks to make feature abstractions, such complex combinations of semantic objects are underpinned by different mathematical machinery [52].

In the middle of the framework is the computational layer based on that found in Marr (1982), which is focused on accomplishing the tasks set out at the semantic layer [35]. These may include classifying sensory inputs or modulating and generating signals required by supervised learning, which are functions computed using items such as n-bit digital numbers [1], neural spikes [38], or analog signals [27]. This layer is largely composed of neural networks, which are often described in terms of layered neurons and their degree of connection with one another. The aim of these systems is to make accurate classifications, stabilize local circuitry, and train efficiently [52]. Various architectures are used to learn in this layer and progress toward these aims, and these computations are usually structured around gradient descent optimization.

Next, the functional layer refers to the fundamental implementation blocks

that perform the base functions through logic gates, artificial neurons, and other circuit components which process signals and shape activation functions [52]. Most importantly to the design of the learning processes at this layer is the creation of effective learning rules for neurons, whether in spiking or non-spiking networks.

Finally, the physical layer encompasses the physical processes involved in implementing all the components mentioned. Specifically, transistors have accomplished much of the physical computation in traditional computers, but research on quantum computing and other forms of implementation such as wetware processing are intended to fundamentally change the nature of this AI hierarchy from the ground upward [52].

Before continuing with similar frameworks to this one, I will briefly discuss the interrelations between these layers for the sake of outlining how the capabilities of AI systems are shaped, especially insofar as the levels themselves contribute to developing the field.

3.2.3 Tradeoffs and Anticipating the Usefulness of A Continuous Level System

The primary focus of this five part system when discussing the features of the layers themselves is how they individually face complexity-performance tradeoffs, or how together the layers face a control-complexity tradeoff [52]. In alignment with the nested LoAs mentioned earlier, this framework views the field of AI hierarchically, with each level serving a particular function with the perspective it conveys, and with consecutive layers providing increasingly detailed information on the model in question.

To begin understanding how the interactions between different LoAs influence our interpretation of a system's complexity, we can examine some tradeoffs identified within this five-level framework. These five notable interrelationships set certain constraints on the AI system's function and performance.

Initially, we might examine the relationship between power efficiency and complexity. As some high-level, complex capability is brought to bear on the lower levels computing it, efficiency often improves. For example, hardware acceleration is useful primarily because implementing a neuron directly on hardware is more energy efficient than doing the same in software [52]. Conversely, the complexity of an AI functionality increases along with its power usage while moving up the hierarchy to achieve more abstract reasoning over sophisticated semantic objects.

Furthermore, there is a notable relationship between transparency and reliability insofar as higher level components are typically more amenable to an intuitive human understanding. It is often easier to inspect and debug the semantic objects which are more readily analyzable. On the other hand, lower level components like neurons and weight matrices are more challenging to query and their internal states can be difficult to precisely characterize. This difficulty can require special tools and methods of analysis, which makes them less reliable.

Likewise, high level components are typically more flexible, since behaviours

can be altered via their more interpretable semantic objects and symbolic representations [52]. As we descend to lower levels however, control over the specific system operations can be more difficult, since manipulating the same behavior can require an understanding of each physical signal involved in the hardware.

In particular, the relationship between the functional and physical layers directly impacts hardware design. Is the structure of neural networks become more ubiquitous, hardware designers can cater to their specific needs by designing architecture that optimize for these designs [52]. For example, the performance of complicated patterns of connection between neurons in neural networks can be accelerated using shared memory blocks. Further, the advancements previously mentioned in quantum computing and other relevant domains promise to continually increase the capabilities of the building blocks of computation.

To connect this to our broader discussion of the nature of discrete leveling systems such as Marr's in comparison to LoAs, we may briefly examine the work of David Danks for insight into core issues underlying these tradeoffs [14]. For the present purposes, it is worth sketching two of his criticisms to anticipate the usefulness of more fine-grained approaches than that of Marr, especially since these criticisms intersect with LoAs in a clear manner. More specifically, these two criticisms of distinct levels of representation in the cognitive sciences such as Marr's are a varying notion of realism and a lack of precision when considering intertheoretic relations.

On the first point about realism, Danks says that this dimension assesses the relationship between a theory and the real world phenomena it describes. Although effectively all cognitive theories have a degree of realist commitment there is variation among precisely which aspects of their analyzed systems are considered real [14]. Especially in the case of using metaphors for higher level concepts that do not literally exist, the boundaries of these metaphors can become unhelpful or misleading. To translate this to LoA terms we can recall that the process of definition makes our ontological commitments clarified including those with regard to realism.

On the second point about intertheoretic relations, Danks describes how there is a predominant tradition where higher level theories are expected to be explained by or reduced to those at a lower level [14]. In place of this reduction process, he posits a notion of constraint between theories which is intended to encapsulate when a change in understanding in one theory leads to change in another [14]. To again indicate toward translating this into LoA terms, we can consider this constraining process similar to the moderating done upon either a LoA to produce its behaviour.

Thus, although it is clear that Marr's levels and its descendant frameworks provide a novel shorthand approach to analyzing complex systems, we can see from these innate tradeoffs that there remains room for improvement. Before articulating that improvement further, it is worth seeing an example of how normativity is intended to be derived from the use of such a leveling system.

3.3 Beyond Marr's Levels for Normative XAI

As has been established in previous sections, both the original three-level and other expanded versions of Marr's framework make for a generally acceptable method for decomposing and understanding complex systems[12].

To illustrate its usefulness, we can first recall that Marr's computational level of analysis is intended to provide explanations for why the computed function is appropriate for its designated task [53]. This is rooted in the nature of the questions asked by this level of analysis; namely, those surrounding the overall trajectory of the function being computed and its integration within the environmental context in which it is situated. As such, the nature of the explanations being provided at various levels of analysis allow for comparisons based on the criteria of the imagined users asking the questions associated with each level. Of course, the plurality of potential explanatory criteria does not necessitate a stable basis for comparison. That is to say, various users may seek explanations for various aspects of the machine learning process, so there must be some means of associating the explanations to the criteria they fulfill. Otherwise, a user's explanatory criteria are left unanswered. This begins to establish how the lens of Marr's levels translates between the realms of understanding and normative explanation, especially since the normative domain of comparisons of explanatory strength are central.

We can now describe in more detail how normative criteria are extracted from users for the purpose of making opaque computing systems transparent. This process will then be translated to suit a LoA-based method rather than Marr's framework in the following subsection. Specifically, the next subsection will connect the aforementioned account of AI opacity impeding upon our understanding of a model's inner workings with the natural alignment of understanding by means of LoA-based method.

3.3.1 Normativity through Marr

Thus far, I have striven to articulate the value of a general leveling framework (of which Marr's represents one specific form) toward understanding advanced ML systems. To give a more precise example of how we may derive normative requirements from such a framework, we may turn to work of Carlos Zednik (2021) [63]. In alignment with Zednik's analysis, our present goal in extracting normative requirements for explanations is to structure these explanations as comparable in their achievement. Alternatively, the comparison process among a multitude of explanations is performed by reference to such norms. For an example in Zednik's terms, we can examine the success or failure of an explanation toward answering a certain why or how question, since these are the norms he has identified in the form of the stakeholder questions.

[12]The same conclusion applies to a LoA-based method insofar as the structured decomposition by way of Marr's levels can be accomplished in LoA terms, though the precise characterization in LoA terms may not be reversely translatable to Marr's levels. Further expansion upon this point will come in Section 3.4.

To briefly overview his translation process, Zednik's normative framework describes a rough mapping from Marr's levels to normative criteria to assign the appropriate explainability methods to each user's concerns. Despite encountering a more detailed description of the Forms of Opacity in Section 2.2.1, we can briefly revisit this issue for the purpose of understanding how Marr's framework intends to dispel the relevant form. Namely, Zednik acknowledges the plurality of opacity issues within ML and states that this creates a series of black boxes depending on the user involved. One central claim he makes is that there is no such thing as one unified issue of opacity insofar as a model is equally inscrutable to all stakeholders. Instead, there are various opacity problems presented to the different stakeholders who interact with the model for which we seek an explanation. As a result, each ML system constitutes a black box insofar as its workings are not particularly amenable to any one user's understanding, even by the experts involved in their creation. It is this multitude of opacity issues which motivates the central concept from Section 2.2.2; The Relativity of Explanation. Simply put, the act of confronting the black-box nature of a system and seeking an explanation cannot be done without a stakeholder standpoint from which to begin seeking. Every stakeholder brings a particular set of concerns to bear on the system in question which manifest as various explanatory requirements.

Toward this end, Zednik would likely agree with our previous analysis insofar as he proposes that opacity must be understood as an agent-relative phenomenon [63]. In contrast however, his conclusion is borne out of an acceptance of an analysis of opacity in computing systems made by Paul Humphreys (2009) [26] rather than something akin to our previous discussion involving the taxonomy of opacity. Instead, this new vision describes computing systems as "opaque relative to a cognitive agent X at time t just in case X does not know at t all of the epistemically relevant elements of the system." ([26], pg. 618). First, we must note the agent relativity of this account which aligns with the Relativity of Explanation. Second and in alignment with our discussion of the taxonomy of it, opacity is an epistemic property concerning a lack of a certain kind of knowledge [63].

This first means that various user agents encounter opacity differently in their interactions with the system based on their designated roles. For example, those who operate an AI system likely consider different aspects of the system's functioning than those who are subject to its decisions, and similarly those creating the AI systems will be concerned with different aspects from those tasked with examining its compliance and safety protocols [63]. In a manner which should sound familiar given the Relativity of Explanation, these operators, decision subjects, creators, and examiners all inquire differently about the functioning of the same system.

Due to the nature of these different inquiries, the characteristics defining each user in their participatory role correspondingly constrain the scope of relevant questions they may ask about the system. Thus, Zednik argues that each predefined type of stakeholder can be matched with the set of questions that are relevant to their standpoint insofar as explanations for them would provide

answers relevant to their explanatory criteria. More specifically, these explanatory criteria are found within what, why, how, and where questions, such as asking why the system made one categorization over another, or how the system achieved a certain result [63]. Of course, these questions may not take on exactly this form, but the overall sorting of concerns into similarly defined categories is more honest to his intention than the precise questions being asked.

On the second point about epistemically relevant elements (EREs), we can now expand upon their nature to articulate the role they play in understanding. According to Zednik, we can proceed from this notion of agent-relative opacity and ranges of possible questions to identify the EREs which help construct meaningful answers [63]. More simply, these EREs constitute the building blocks of the knowledge useful for reducing a system's opacity [26]. We can thus expand upon this notion for this context by specifying that an element may be "some step in the process of transforming inputs to outputs, or as a momentary state transition within the system's overall evolution over time" ([63], pg. 269). For such an element to become epistemically relevant, it must be known and also capable of being cited for the purpose of explaining another element's behaviour or some aspect of the system's output [63]. In this context, the process of explanation becomes the acquisition of knowledge of a systems EREs, though there remain degrees of freedom by which to deliver various explanations. Zednik says that whether we explain computation electronically through the magnetization of hardware registers, mathematically using the manipulation of binary strings, or rationally by reference to the system's goals and representational states, these explanations remain equally legitimate ([63], pg. 269). However, in alignment with the Relativity of Explanation as posited herein, these various explanations serve particular stakeholders depending on the task they wish to complete.

To amalgamate one of Zednik's examples with one made previously about interacting with a virtual assistant, imagine such an assistant being used to make a small online shop by integrating with existing e-commerce tools. It is easily foreseeable that if the internal representation of some useful data such as a date given by a verbal prompt to the virtual assistant could be in a format that requires slight modification to integrate with a third-party tool[13]. From our earlier engineer example perspective, this means that correctly operating the system would depend on this ERE being clarified such that it adds a small amount of transparency.

For the purpose of translating these notions into terms consistent with the rest of this thesis, we can note that agent-relative opacity can be interpreted as fundamentally akin to the Relativity of Explanation. As will be further demonstrated in Section 3.4, the constraints delimiting the range of possible questions are intimately tied to the choice of LoA. Before this demonstration however, we can quickly follow Zednik's argumentation to show how the alignment of each user type with its set of questions is achieved using the guardrails of Marr's

[13]It may be noteworthy that the steps involved in this online storefront generation process may be found in more detail than is necessary for present purposes [46]. This is simply to say that such online integration is not a purely theoretical example, and we could imagine storing dates as either DDMMYYYY or MMDDYYYY, and this causing problems downstream.

levels.

The first notable item allowing Marr's levels to be used as guardrails is that we may divide the potential stakeholder questions according to their most relevant level. Specifically, concerns about what a system is doing and why are handled at the computational level [53], questions about how the system performs its actions are addressed at the algorithmic level [37], and questions about where the relevant operations are realized falls to the implementational level [62].

As such, Zednik argues that the explanation-seeking questions posed by each distinct level funnel the identified user requirements toward those EREs fulfilled by existing techniques [63]. Then, insofar as the identified XAI technique fulfills the relevant explanatory requirements, Zednik asserts that the user's explanatory requirements have been successfully met [63].

We can briefly examine one noteworthy example corresponding to each level to clarify this process. The algorithmic example will be elaborated upon in the next subsection to discuss the generality of LoAs in comparison to Marr's levels. Presently however, Zednik proposes that one technique which answers questions about what a system is doing and why it produces certain outputs is done by input heat mapping [63]. In the most abstract, computational-level sense, this technique involves highlighting features of the input which bear the responsibility for specific outputs according to the algorithm. In terms of answering how questions for the algorithmic level, we can instead use feature-detector visualization to identify system variables which detect and characterize the specific representations of input features and their influence on the overall system behavior. A notable example of this can be found in one dissection of generative adversarial networks which aimed to identify the units most sensitive to uniquely recognizable features in the input (crosses on buildings for identifying churches for example)[6]. Finally, implementational level questions correspond to the method of diagnostic classification when no clear feature detectors are present. This technique determines which information becomes represented by a system after receiving some input.

As a result of Zednik's analysis, we have now seen how the formulation of stakeholder-specific questions leads to a more tailored normative description of which style of ERE corresponds to which existing explainability method. In addition to the agent-relative notion of opacity which motivated the need for plural explanation methods, we have also been exposed to other argumentative features marking this account as normative which are worth clarifying before continuing. Specifically, we have seen a means of evaluating the explanatory success of a given explainability method by reference to the norm of its ability to provide EREs to Marr-level questions. To conclude with a further metacognitive point, the gaps surrounding currently available explainability methods demonstrate some of the limitations associated with the techniques to indicate what types of explanation cannot yet be achieved.

With this overview of Zednik's approach to normative explainable AI using Marr's levels, we can now progress to a discussion about the advantages in generality made possible with LoAs as opposed to Marr's more coarse-grained

levels.

3.3.2 Generality of LoAs for Normativity

As has been stated, there are myriad ways of explaining the behaviour of a computing system depending on the EREs sought. We have now seen how Zednik provides a breakdown of explanatory pathways which navigate through the filter of Marr's levels to determine which available explainable AI method can most aptly handle the concerns raised by a particular stakeholder's question. As such, this account makes progress toward creating a normative framework evaluating the successes of explainable AI methods. However, in accordance with the conclusions drawn by Hamrick and Mohamed [22], by Niv and Langdon [40], and by Zednik himself [63], there appears to be a clear methodological need for a more fine-grained framework for analysis which, at the very least, allows for more flexibility and modularity. Some further consequences of this movement away from such a coarse-grained framework toward a more fine-grained or continuous one will be elaborated in the next section.

In the least sophisticated sense, LoAs are clearly more general and flexible than Marr's levels since LoAs can take arbitrary observables as their constituents. We have already been exposed to this continuity as a gradient of abstractions (GoA). This formulates LoAs as a continuum rather than as discrete levels within Marr's framework, which brings inherent flexibility in its approach to modeling.

Moreover, we have seen that Marr's levels may be subject to criticisms surrounding its presumption of altered ontological commitments, but this does not hold in the case of LoAs. Instead, LoAs attempt to understand and explain the target system while clearly delineating how the perspective taken is altered. This attempts to equalize the realist commitments made across LoAs [20], insofar as none are presumed to be more real than any other.

To briefly exemplify the precision of LoAs in comparison to Marr's levels, we can review in more detail the algorithmic-level questions and explainability method alignment previously provided in Zednik's normative account. In short, the algorithmic level focuses on how a system accomplishes its specified functions, and answering "how" questions at this level bridges between the tasks of computing the correct functions for the designated task, and correctly computing these functions.

In LoA terms, this algorithmic bridging is related to the distinction between external adequacy and internal coherence, between which we may choose a GoA that is "realistic" ([20], pg. 325). Specifically, external adequacy means the LoA chosen adequately reflects the phenomena being studied and thus is computing the correct functions, whereas internal coherence implies that the chosen LoAs in the GoA should be logically consistent and should not result in contradictions with one another.

Further, it is helpful to remember the distinction between disjoint and nested LoAs. This is especially useful when determining how "parallel" processes (when viewed from a hierarchical stance) could influence one another. Especially in

these cases where there is some overlap between LoAs, it could become impossible to disentangle potentially independent systems using Marr's levels.

If we recall the example of a feature detector analyzed at the algorithmic level, suppose we imagine such a system embedded in a self-driving car. There are two related features of such an example which support the notion of LoAs being more flexibly applicable and maneuverable between these constraints than Marr's algorithmic level: explicit goal comparison and hierarchical precision. Marr's framework is useful when describing specific algorithms but can only weakly compare those at different levels or further, validate their consistency. Moreover, there is the possibility for an explicit inclusion of higher-level goals in a hierarchical analysis with LoAs which is not present for Marr. For more detail on the difference of mechanics of these processes, we will elaborate on how this understanding is made clearer in the next section. For the time being, we can briefly demonstrate how these might be accomplished in the vision system of a self-driving car.

For the example of explicit goal comparison in a self-driving car, the explicitly defined LoAs can fit the exact contours of the self-driving system such as for sensory perception, decision-making, and vehicle control. So, LoAs can more easily reveal how one level in sensory perception (object detection) interacts with one for route planning (decision-making). The bidirectional relationship between the two would be more easily characterized as each route decision caused different sensory perceptions, especially in a situation where both are happening in rapid succession such as crash avoidance.

To touch on the example of hierarchical analysis in the same system, LoA definition can more straightforwardly assist with both top-down and bottom-up analyses as the various subsystems interact across level boundaries. Likewise, this can help identify feedback loops and previously unknown dependencies; perhaps a change in the vehicle control processes (a tire going flat, unexpected conditions, or impact) could influence whether the car maintains the ability to follow the planned route.

We can now speak in broader terms about the contributions to our understanding provided by LoAs, especially in the case of AI.

3.4 Understanding AI with Levels of Abstraction

As we have begun to see with the self-driving car example, the relevance of these features provides insight into the resultant relationships that form between levels. Thus far, I have argued for the advantages of LoAs over Marr's levels. After discussing multiple authors who - despite appreciating the value of level-based frameworks - support the need for a more continuous approach than that offered by Marr, I aimed to show that one attempt at normativity through Marr's levels can be made more precise due to the more generalizable capacity of LoAs. For the present purposes, I shall now elaborate upon the mechanics of how LoA-induced perspectival shifts affect our understanding of their target system, and their role in elucidating a model's function. As discussed, the choice of variables and their behaviorally-constrained values provide a selection

of possible relationships between observable features of the system and the LoAs they constitute.

We can now recall how Khalifa's account of comparing degrees of understanding between subjects enables normativity to be established. More specifically, the most relevant of his conditions for understanding required an account of how the relations between the explanations establish a meaningful form of understanding. As such, we will now see how the capacity to analyze these relationships between LoAs with precision forms the basis of our understanding explanations, as well as providing grounds for evaluating their success through this notion of comparison.

Where Zednik began his sorting process by differentiating stakeholder requirements by the type of questions they would plausibly ask, I will instead argue that the structure of a LoA is inherent within the notion of an Explanatory Nexus. Then, I will expand upon some of the potential strategies for novel explanation that arise when using a system of LoAs, thereby furnishing our Nexus. I finish this section by situating this approach to understanding AI within some relevant literature.

3.4.1 The LoA Structure of an Explanatory Nexus

I will now posit that Khalifa's notion of an Explanatory Nexus can be helpfully structured in LoA terms. As such, a Nexus offers a fitting method for establishing comparison between competing explanations due to the advantages inherent in the LoA approach overviewed previously. However, this is not meant to straightforwardly equate an explanation with a LoA. Instead, we can dissect the component pieces of an Explanatory Nexus and show how each benefits from the LoA structure as mentioned in the previous section. This will further demonstrate the flexibility in this approach which is not present with Marr's three levels.

At its core, the Nexus functions as a connection between our understanding of a phenomenon and its explanation to capture the relationship between what is known and how it is known. More specific to its structural qualities, the Explanatory Nexus is the set of correct explanations of some proposition as well as the relations between those explanations ([28], pg. 6). The inter-explanatory relationships between these supporting propositions ground this Nexus within a network of related arguments. The EKS model then establishes that understanding is simply the process of a subject coming to grasp this Nexus and for it to bear resemblance to scientific knowledge as described by the Scientific Knowledge Principle ([28], pg. 14). If we can say that the subject's Nexus bears some degree of resemblance to approximately true scientific knowledge, they have a minimal degree of understanding (as opposed to none) [28]. As discussed in our explanation of the EKS Model, these features beget a spectrum of understanding from the minimal to the more richly connected Nexus in the ideal case [28]. This makes the EKS model a useful tool for comparison and helps to connect the realms of knowledge and explanation. With the branching network image of an Explanatory Nexus in mind, we can proceed to an illustration of its core

features in LoA terms.

More precisely, both the explanandums (statements to be explained) and explanans (relevant explanations) involved in the construction of a Nexus can be formatted in accordance with the LoA method, thereby granting the benefits discussed previously. For the sake of making worthwhile comparisons across LoAs, I have chosen examples which vary the LoA while referring to the same type of explanandum/explanans rather than simply providing more detail. However, the notion of providing more detail will be captured by briefly mentioning the effect of distinction between nested and disjoint LoAs in this context. With this distinction, the habit of providing three LoA examples will be differentiated from the coincidental overlap with Marr's three-leveled framework.

For the explanandum side, a potential observable at a low LoA could be a classifier's decision instance, where the input and output might be the pixel values of the image and the associated label. Moving the LoA upward, an intermediate observable could be the patterns across multiple classifier decision instances, where the inputs could be common aggregate features from a set of images and the outputs would be the trends in consistent misclassification. Finally at an even higher level, one further observable could be the general behavior guiding a classifier's decision making, where the input is now the entire architecture including its parameter values and the output is the general behavior it exhibits.

And for the explanans, one potential observable for a low LoA could be which specific factors influenced a particular classification instance, where the input would be the features detected in the image and the output being the resultant classification. Moving upward again, a mid-level observable for explaining patterns of decisions could be the choice of algorithm, where the input would be an images features and the algorithm used to make the classification with the output being the associated waiting given to those features by the various components of the algorithm. And finally at the highest example LoA, an observable might be a foundational algorithm whose input is a description of the error gradient for some training instance, and the output is the parameter weights adjusted after optimization.

As we can see, both the explanandum and explanans can be afforded various LoAs. Although it may appear as though Marr's levels could do the same task, it is important to recall the arbitrary level of precision that can be sought within this LoA framework which, admittedly, may align with Marr's levels at times. Now in the case of the explanans, which presumably affords more interesting results when varying the LoA chosen for the same explanandum, we can of course further decompose these explanations to contour their strengths around the needs of various users.

For this decomposition, it can be helpful to recall the distinction between nested and disjoint LoAs, which further distinguishes this method from Marr's style. After outlining some of the potential benefits of this distinction, an argument gesturing toward further steps in this contouring process will be conducted in Section 4.2. Essentially, a nested GoA is primarily helpful in describing complex systems precisely with LoAs which become incrementally more accurate in

their modeling of the system in question ([20], pg. 313). In contrast, a disjoint GoA describes a system as a combination of non overlapping LoAs ([20], pg. 313). This case is relatively simple since oftentimes the LoAs are overlapping in some of their observables.

Nevertheless, this distinction reminds us of the benefits associated with clarifying the nature of the LoAs being used. More specifically, nested levels provide a cumulative insight as each layer builds upon the others, where moving through layers allows one's understanding to become increasingly comprehensive. Moreover and as indicated previously, nesting affords a continuity to the analysis as each level fits like a puzzle piece into the last. This helps to create a cohesive narrative from the bottom to the top of the system, so to speak.

In contrast, which may offer a unique angle on the problem, especially if the problem requires focusing on a specific LoA without considering others. In instances where higher dimensional objects are easily conceptualized it could also be the case that these split levels make for a clearer encapsulation of the target system[14]. Where nested levels bring coherence, disjoint levels may offer unique insights especially in contexts where functionality in isolation is helpful such as troubleshooting or refining AI behaviours.

Now, we can examine some examples which clarify the type of continuous gradation possible within the observables found in an LoA that are not possible within Marr's framework. Of course it is foolish to attempt enumerating a list in a continuous space, but for the sake of showing the fine-grainedness of these observables in contrast to Marr's levels, we can briefly list a few items. Where iteratively applying Marr's framework would restrict the possible views of a given entity to three perspectives at a time, we can instantiate arbitrarily many. In the case of nested LoAs, a more continuous spread of neural network weights in a decision making process could range from the individual parameter weights at the low end, to the activation of individual nodes in response to input, to the activation patterns of a whole layer to the combined decision making logic from the amalgam of all layers, to the end classification in a given context at the higher end. In the case of disjoint levels, an example of a more continuous spread is admittedly difficult to conjure since we would need some problem where the number of break points between LoA is unrelated to understanding the overall system. Nevertheless, sound waves being captured in compartmentalized maybe one example. In one disjoint set, we could have the raw sound waves as captured, or this sound could be broken into its phonetic components, and finally another arrangement could work at the level of recognizing words from these phonetic components. In each case, the specific LoA for each component could be seen as independent from the rest.

So what benefits can be drawn from understanding the relationships between LoAs within this notion of an Explanatory Nexus? I now briefly identify three which will be explained in the next section. These three benefits are, at least theoretically, an increased holistic understanding of the system, the ability to

[14]Among the systems which may benefit from this type of analysis is the earlier description of ranks- and files-chess, which is effectively an n-dimensional array representation of a given board state.

identify root causes of behaviours, and the bridging between stakeholders with various degrees of technical experience.

3.4.2 Novel Explanatory Strategies

The aim of this section is to briefly summarize some of the overarching themes in the usefulness of the LoA structure as I have described it as being present in an Explanatory Nexus. The first three identified were an increased holistic understanding of the system, the ability to identify root causes of behaviours, and the bridging between stakeholders with various degrees of technical experience.

In particular, an improvement in one's holistic understanding of a system refers to the manner in which we can move between levels to get a more comprehensive view. Moving from the lowest level - where we see the most immediate reasons for specific outcomes - toward higher levels which encapsulate further-reaching patterns, our explanations portray a broader understanding of the system.

In the case of root cause identification, there may be instances where lower-level anomalies can be traced back to foundational aspects or assumptions within the algorithms. Navigating between these levels can illuminate issues related to their integration with one another, especially regarding the relations between levels when explicated in LoA terms. A helpful example of this root cause clarification process comes in the form of the ontological commitments borne out of defining a LoA. Examining the alignment between the ontological commitments made at one level and how they relate to those made at another may illuminate an incongruence which could be producing a root issue[15].

And lastly in the case most relevant to this thesis, we have seen an example of how various stakeholders are likely interested in different LoAs since they afford different forms of understanding within an Explanatory Nexus. In Zednik's terms, we can envision the needs of operators, decision subjects, creators, and examiners as different but potentially overlapping with one another. However, where Zednik uses Marr's levels to cluster explanations together in somewhat broad, inflexible terms, the LoA approach as embedded within our Explanatory Nexus offers more holistic integration across levels, ontological clarity, and the opportunity to connect the most relevant explanations to an arbitrary stakeholder's requirements.

Aside from only bridging between stakeholders with various degrees of technical experience, we can also describe how bridges between levels can afford their integration and offer novel explanations. We can recall that Section 3.2.2 discussed the properties of explicit goal comparison and hierarchical precision for the purpose of showing how LoA offer the benefits of a more general framework than Marr's levels. Now however, we can discuss the possibility for relations defined between LoAs to provide a means of rigorously generating bridge laws.

[15]This is made particularly clear in Danks' analysis of the blurring between ontological commitments at different Marr levels [14]. Especially insofar as cognitive theories are analogized to AI processes, we must remain wary of any alterations in the commitments to realism made between levels.

The focus of this description will be to merely support the existence of such novel explanatory strategies, without articulating the extent of their capability. A further speculation for the potential usefulness of such a method will be provided in Section 4.2.

In her attempt to articulate the space for novel explanation using the concept of abstraction, Eleanor Knox aims to demonstrate that the process of description which moves from a theory in one LoA to another can lead to explanatory novelty ([29], pg. 44)[16]. To do so, she derives novel explanatory strategies originating in the physical transition from thermodynamics to statistical mechanics via complex bridge laws.

For the purposes of our discussion, the realm of statistical mechanics provides precise detail to focus on the most minute interactions in a system. Conversely, thermodynamics is more macroscopic in its description of abstract system behaviors. Where statistical mechanics may encapsulate the precise nature of particle interactions, thermodynamics refers to these intricacies with simpler abstractions such as heat and entropy [29]. The goal of a bridge law then, is to provide a rigorous mathematical framework to avoid fitting too closely to the detail available at a microscopic state of a system and thus being unable to derive overarching patterns, while retaining the information provided at a lower level [29].

Despite using the realm of physics as her arena, this notion of a bridge law allows for a mode of transition between LoAs which is applicable to the domain of AI. For example in the realm of reinforcement learning, we could construct a bridge between the algorithm's state-action reward dynamics to a simplified higher-level LoA which clarifies aspects of the decision-making process. More specifically, bridge laws could help elucidate the relationships found within not only the data themselves, but perhaps also between the other lower-level features of the algorithm itself. Such lower-level features may include parameter values, the nature of the layers involved in a classification, their constituent nodes, and the connections between them. As is consistent with this account of bridge laws, understanding and intervening upon the behavior of the overall algorithm at this level may be effectively impossible due to its nature as highly-dimensional and complex.

This sounds reasonable enough in these generic terms, so perhaps it is necessary to provide more description of the specific form such bridge laws may take. Of course, the task of specifying the relation between LoAs is clearly not a straightforward one. For this purpose, it may be most appropriate to design bridge laws consisting of rule sets or algorithms which connect higher-level representations of data (like clusters, categories and features) with the low-level raw data (data instances or pixel values). In a somewhat familiar case of a visual feature detector, this bridge law could clarify the logic mapping from pixel

[16]For brevity, I have taken the liberty of substituting Knox's broader notion of abstraction for an explicit usage of the LoA terminology. This follows relatively clearly from applying our understanding as formatted within the LoA-structured Explanatory Nexus, but it is worth noting she does not make use of LoAs verbatim.

values to object recognition, rather than performing the recognition itself[17].

Nevertheless, we have seen how LoAs offer an opportunity for novel explanation due to their simplification of information shrouding the relevant conclusions from being drawn. Knox identifies the potential for these explanations by saying "changes of descriptive quantity induced by sufficiently complex bridge laws lead to new standards of abstraction, and thus novel explanatory strategies" ([29], pg. 57). As such, it is worth investigating further how LoA-facilitated bridge laws could be generated to describe the relations between the vast arrays of individual components and their interactions en mass. Although the form of such bridge laws has clearly not been made precise in this section, it seems to be an appealing avenue for further work in the domain of explanation due, at least in part, to its apparent success in the transition between statistical mechanics and thermodynamics.

3.4.3 Situating Understanding through LoAs

We can now situate the usefulness the aforementioned method of analysis in relation to the taxonomy overviewed in Section 2.2 to show how its scope has been expanded from that proposed by the original authors of the taxonomy [19]. Specifically, this method of analysis is the embedding of an LoA structure within an Explanatory Nexus as a means of clarifying the relations between explanandum and explanans.

Within the taxonomy that we have examined, LoAs have already been identified as centrally useful within access opacity. However, it is my contention that this form of analysis described herein is also useful beyond this category. More specifically, consider trying to understand the nature of link opacity and a stakeholder's explanatory requirements. If these requirements remain unfulfilled and thus link opacity holds, this means that the empirical link between the core concept of a model and its relation to the target phenomenon in the world is weak [55].

However, this neglects any mention of the relevant LoA used for determining which target phenomenon in the world is matching up with exactly which elements of the model. More specifically, it is necessary to determine an LoA such that a reasonable comparison can be made between this notion of link opacity with its empirical support and those aspects of the model which are deemed worthy of matching with the target phenomenon[18]. If an analogy is to hold

[17]This distinction between performing the computation and clarifying the logic performed within it is relevant to the role played by such bridge laws. However, further elaboration on this point would only constitute further speculation beyond the scope of this paper. Regardless, it is worth noting that the use of LoAs in these cases would need to be subject to an iterative application of constraints from the external adequacy (validity) and internal coherence (verification) to facilitate the transition. The interactions between the LoA (as it is defined for a stakeholder) as it fits between these two constraints may offer some avenues for articulating the specific role that a given ERE is playing in the wider system function.

[18]Within a similar taxonomy for understanding in the domain of AI [48], a version of link opacity is seen as a form of understanding without interpretability. If we maintain that the concept of link uncertainty rests at the heart of such a form of understanding, a similar argument would apply. Specifically, if "the degree to which the central concept captured by

between the function of a some model of a world system and its supposedly analogous mechanism in the world, we must assert some grounding for that analogy. Otherwise, even the notion of providing empirical support for a model functioning as this world-system analogy would be without a target. Perhaps more clearly, we cannot provide empirical support for any internal aspect of a model without deciding that such empirical support take a certain form, a form which is dictated by an implicit LoA choice. As a result, there is an implicit LoA choice presumed by the notion of link opacity which permits coherent comparisons.

Although it is notable that this choice of LoA is not identical to the options proposed within access opacity[19], it should at least be clear that the domain for an LoA based method of analysis extends beyond the reach of only access opacity.

4 Domains of Application

This final section will begin articulating a few case examples of the applicability of LoAs in the analysis of AI systems more broadly. As such, it is the briefest section in this thesis due to its comparatively speculative nature.

Some authors believe that LoAs are already widely accepted in the domain of computer science [19], and perhaps in some regard this is true of computation in general. However, it seems as though their usefulness when analyzing advanced systems such as those used in the domain of AI has been minimally explored as an explicit tool. It is for this reason that I use this short section to gesture toward some potential directions for further application of this method.

For the sake of reconvening on a notable topic mentioned in the introduction, I will sketch an argument for how the use of LoAs can clarify certain claims made about the responsibility gap. In particular, I will argue against one such paper which concludes that the gap does not exist (Section 4.1). Further, I indicate how LoA are currently implicitly involved in describing sets of normative criteria in modern research, as well as how bridge laws structured by means of LoAs might assist with the plight of modern interpretability of neural networks; one of the more prominent issues within the domain of AI safety (Section 4.2).

4.1 Interpreting Responsibility Gaps in Machine Learning

Now, we can briefly apply a LoA-style analysis to one paper arguing against the existence of responsibility gaps. As such, the aim of this section is to circle back provide some degree of completeness to this aspect of the literature review and offer an example of the usefulness of LoAs as a tool for thought, rather than formal analysis.

a model maps to the target phenomenon in the world, as measured by the empirical support and linkage to the target" ([55]), then the following argument in favor of broadening the scope of LoA analysis beyond only being mentioned within access opacity (as in [19]) is supported.

To begin, we can recall that the responsibility gap essentially refers to the space that exists between the increasing capabilities of autonomous decision-making and the lack of corresponding responsibility-attribution methods to account for this expanding class of decisions. To recall the description offered by the author who coined the concept, we must remain wary of the "increasing class of machine actions [that are] incompatible with traditional means of responsibility ascription because nobody has meaningful control" [36]. Daniel Tigard argues that this problem doesn't exist, however, as long as we have a clear understanding of the values embedded within a technology and can acknowledge the multitude of existent practices for associating actions with responsibility [56].

For the sake of tying up this responsibility gap loose end and demonstrating the informal use of an LoA analysis, we can outline the most relevant pieces of his argument against the existence of a responsibility gap and identify two interconnected reasons as to why his conclusion is weak.

Initially, Tigard argues that there is a plural ambivalence in our current responsibility attributions. This means that we often hold individuals or broader societal entities responsible in a variety of ways depending on the context, incident and type of responsibility involved [56]. Among other methods, this also includes our attempts "to locate accountability in technology", which allows us to "engage in other forms of responsibility practices" (both quotes from [56], pg. 599). He continues by arguing that as with other forms of responsibility, responsibility can likewise be managed in the context of complex technology. Moreover, despite their growing autonomy, lack of humanlike intentions and capacity for moral agency, these technologies can still be manipulated and contained [56]. Finally, Tigard acknowledges that those impacted by technological decisions will likely seek to understand the embedded underlying values within the systems which, presumably, we were unable to properly manipulate and contain. He then concludes that understanding the nature of these values and their role in the design cycle facilitates the assignment of responsibility to designers, users, or even the technology itself [56]. With this brief summary of some of his core points in mind, we can proceed to a description of the two interlocking issues.

The first issue Tigard encounters is in his oversimplification of ethical scenarios. This important to observe that the core of this argument rests upon how understanding technologically-embedded values effectively clarify responsibility. Unfortunately however, this simplification offers little by way of deeply understanding the intricate ethical landscape surrounding and influencing advanced autonomous technologies. This oversimplification can be seen as a misappropriation of an LoA, insofar as it does not adequately reflect the proper observables for tracking the impacts of values across the development and deployment of an autonomous system. More straightforwardly stated, this referencing of "understanding values" does not provide any basis for responsibility attribution that goes beyond the issues of opacity we have already encountered. At such, there remains ambiguity in determining the bearers of responsibility for the outcomes affecting concerned stakeholders. This brings us to the second issue which is

very similar and which is closely related to our previous discussion of LoAs.

Specifically, it is this resultant lack of clarity in Tigard's responsibility assignment. Building upon the premise that understanding values clarifies responsibility, it remains difficult to apply to situations where there are inherently many influences on the creation, deployment, and operation of such autonomous technologies. Especially in cases where understanding an autonomous system's capabilities is limited by a degree of access opacity as described in section 2.2, Tigard's account struggles to provide any robust foundation for determining responsibility when these values and underlying processes are themselves obscure. As such, this demonstrates how approaching an argument through the lens of a LoA analysis can clarify our understanding of notions such as the existence of the responsibility gap.

4.2 Generating LoA from Normative Criteria

One final remark is worth making on the comparable value of LoAs versus Marr's levels with regard to their applicability as either a tool for thought or explicit analysis. Namely, there may certainly be instances where the conversational maneuverability of Marr's levels makes them preferable for discussions of certain algorithmic features. However, since only making a useful conversational contribution is not the supposed intention of a systematic analysis for extracting normative criteria, it seems reasonable to maintain the value of the more rigorous LoA approach described herein. For the present purposes of this thesis however, attempting to add notation would only provide a false precision, since I assert that such a project is outside the current scope. Regardless, the distinction between the feasibility and current plausibility of rigorously defining interpretability methods with LoAs this one I wish to maintain.

One especially useful domain where generating explicit LoA to deal with stakeholder relative normative criteria would be in mechanistic interpretability. In other words, mechanistic interpretability refers to the sub-discipline within explainable AI attempting to explain and predict the behaviours of neural networks through understanding the algorithms which underlie these models [42]. One of the central issues in the field of mechanistic interpretability is that of neuronal superposition [18], which is the conundrum where individual neurons participate in the representation of multiple features of the input. As researchers attempt to analyze this problem and provide it with an epistemic foundation at the "microscopic" level [43], we may notice a very similar argumentative structure to that of Knox and the translation from statistical mechanics to thermodynamics [29]. More specifically, there lies an implicit inclusion of LoAs within a core work projecting the direction of future mechanistic interpretability methods [43]. Upon this analogous "microscopic" theory, there lie four proposed methods forward, three of which include clear reference to LoA analyses [43]; larger scale structure, universality, and bridging from the aforementioned microscopic to the macroscopic. These items each deserve another word of clarification after examining the nature of this epistemic foundation.

To that end, the epistemic foundation upon which mechanistic interpretabil-

ity rests is an examination of the smallest pieces of a neural network; namely, its parameter values and features [43]. This LoA has already been identified as one of the lowest practically available in Section 3.4.1. The LoA above this begin to attempt explaining how combinations of these pieces result in individual nodes representing multiple complex aspects of the input.

Here, larger scale structure refers to this "more abstract story" ([43], under Larger Scale Structure) that appears to remain coherent on top of understanding these lower-level features. Is my contention that an explicit LoA analysis would further clarify potential distinctions between these low-level features, if it is true that they are at precisely the same level.

Second, the notion of universal features and circuits posits that neural networks trained on similar domains create similar patterns in their methods of recognition [43]. On the previous discussion of the plausibility of an explanation from Section 2.1.2, this universality may be the closest thing to a ground truth that is discoverable. However, it still seems to be the case that determining which patterns truly are universal (insofar as various AI models really are sharing exactly the same pattern) would be a process simplified by an LoA analysis.

Third, the process of bridging from the microscopic epistemic foundation to the macroscopic patterns is especially relevant for LoA analyses since some microscopic discoveries have already been shown to have macroscopic effects [43], because this is effectively identical to that described by Knox [29]. As such, it seems reasonable to expect that a similar application of complex bridge laws could benefit this pursuit as well[19].

As we have discussed in Section 2.2, the multifaceted black-box problem presents various forms of opacity to the stakeholders involved. As such, the explanatory requirements of such a stakeholder begin to constrain the LoAs which are helpful in their understanding of the relevant internal functions.

Rather than focusing on potential users and the questions they would likely ask with respect to Marr's levels as is Zednik's approach [63], it is perhaps more worthwhile to begin by considering the observables involved in a LoA which provide a sufficient explanation for the system in question, such as a particular neuron in superposition. If we can determine the criteria for a sufficient explanation in this context (given the researcher stakeholder's background knowledge and skills), we could provide measurable targets for successful explanation against which to compare progress toward understanding larger scale structure, universality, or the notion of bridging from the microscopic to the macroscopic. As such, with reference to the process of combining the lowest-level pieces into nodes containing complex representations, we see the value of this arbitrary def-

[19]One notable qualification on this speculation is that Knox concludes that a metaphysical account of emergence and its relation to levelism seems plausible ([29], pg. 57-58), whereas we must recall that the LoA approach only seeks to make claims in the domain of epistemology. Even still, she expresses that her account of explanatory novelty does not fit well on either side of this metaphysical/epistemological debate ([29], pg. 58). As such, further work seems necessary to precisely differentiate the interactions between these novel explanatory techniques and the metaphysical assumptions supporting different accounts of emergence.

inition of explanatory criteria in full [44]. Thus, as a result of the Relativity of Explanation and the continuity of LoAs, we could determine a set of relevant observables that fits essentially arbitrarily defined explanatory criteria to provide normative targets for progress toward these goals in mechanistic interpretability; namely, understanding larger scale structure, universality, and building the bridge between the micro- and macroscopic.

In navigating between somewhat understandable low level features and the emergent properties born out of them which seem entirely unpredictable, we can conclude by advocating once more for the potential usefulness of LoA-structured bridge laws as done previously.

There is much complexity surrounding the notion of emergence across various domains but recently and most presently relevant, in the abilities of language models [60]. This has spurred interest toward directly applying psychological investigation methods to language models [21], though in contrast some authors believe any emergent properties dissolve when interpreted with the correct metrics and thus constitute little more than a mirage of ability [51]. Nevertheless, the creators of some widely used language models have warned against numerous emergent properties that were not explicitly included as intended capabilities [46]. Even with only this short description, it is clear that a full discussion of the concept of emergence is beyond the scope of the current work. As such, it remains a monumental task to attempt understanding the nature of emergent capabilities in AI, and I hope that my proposed LoA-structuring of analysis represents a push in the right direction.

Conclusion

This thesis is has aimed to articulate a clarification of a framework for understanding based on Levels of Abstraction, especially as they pertain to understanding AI models. After a preliminary overview of some of the relevant literature in Section 1, I described an account of understanding which is taken to be a regimented version of the received view in Section 2. I then discussed how obstacles to understanding come in various forms of opacity, and aimed to show how explanation was fundamentally tied to those stakeholders seeking them. Section 3 gave a longer description of LoAs and their structure, and showed examples of similar frameworks being used to clarify normative considerations that adjusted to different stakeholders. I strove to demonstrate the flexibility and precision that may be achievable with an LoA framework in broad terms. I finished Section 3 with a call to action in pursuing LoA-based analyses involving the construction of complex bridge laws which aim to usefully translate between LoAs. Finally, Section 4 briefly mentioned one argument concerning the existence of the responsibility gap and gestured toward further research which aligned with the modern goals of understanding advanced AI systems.

References

[1] F. Akopyan and Filipp. Design and tool flow of IBM's TrueNorth. In *Proceedings of the 2016 International Symposium on Physical Design - ISPD '16*, pages 59–60, 2016.

[2] Ron Atkin. *Multidimensional Man: Can Man Live in Three-Dimensional Space?* Penguin books, 1 edition, September 1982.

[3] Huiren Bai. The epistemology of machine learning. *Filosofija. Sociologija*, 33(1), 2022.

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

[5] C. Van Fraassen Bas. *The Scientific Image*. Oxford University Press, New York, 1980.

[6] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. *ArXiv*, 1811:10597, 2018.

[7] John R. Beaumont and Anthony C. Gatrell. *An Introduction to Q-analysis*. Number no. 34 in Concepts and Techniques in Modern Geography. Geo Abstracts, Norwich, 1982.

[8] William Bechtel and Oron Shagrir. The Non-Redundant Contributions of Marr's Three Levels of Analysis for Explaining Information-Processing Mechanisms. *Topics in Cognitive Science*, 7(2):312–322, April 2015.

[9] Hannah Bleher and Matthias Braun. Diffused responsibility: Attributions of responsibility in the use of AI-driven clinical decision support systems. *AI and Ethics*, 2(4):747–761, November 2022.

[10] Florian J. Boge. Two Dimensions of Opacity and the Deep Learning Predicament. *Minds and Machines*, 32(1):43–75, March 2022.

[11] Nick Bostrom and Eliezer Yudkowsky. The ethics of artificial intelligence. *Artificial Intelligence Safety and Security*, 2014.

[12] Jenna Burrell. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1):205395171562251, June 2016.

[13] Mark Coeckelbergh. Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics*, 26(4):2051–2068, August 2020.

[14] David Danks. Moving from Levels & Reduction to Dimensions & Constraints. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 35, 2013.

[15] Hans de Bruijn, Martijn Warnier, and Marijn Janssen. The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly*, 39(2):101666, 2022.

[16] Roos De Jong. The Retribution-Gap and Responsibility-Loci Related to Robots and Automated Technologies: A Reply to Nyholm. *Science and Engineering Ethics*, 26(2):727–735, April 2020.

[17] Paul B. De Laat. Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability? *Philosophy & Technology*, 31(4):525–541, December 2018.

[18] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition, Sep 2022. Published on Transformer Circuits, affiliated with Anthropic, Harvard, and others.

[19] Alessandro Facchini and Alberto Termine. Towards a taxonomy for the opacity of ai systems. In Vincent C. Müller, editor, *Philosophy and Theory of Artificial Intelligence 2021*, pages 73–89. Springer, April 2022.

[20] Luciano Floridi. The Method of Levels of Abstraction. *Minds and Machines*, 18(3):303–329, September 2008.

[21] Thilo Hagendorff. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods, 2023.

[22] Jessica B Hamrick and Shakir Mohamed. LEVELS OF ANALYSIS FOR MACHINE LEARNING. 2020.

[23] Jan-Hendrik Heinrichs. Artificial Intelligence in Extended Minds: Intrapersonal Diffusion of Responsibility and Legal Multiple Personality. In Birgit Beck and Michael Kühler, editors, *Technology, Anthropology, and Dimensions of Responsibility*, pages 159–176. J.B. Metzler, Stuttgart, 2020.

[24] David H. Helman. Realism and Antirealism in Artificial Intelligence. *The British Journal for the Philosophy of Science*, 38(1):19–26, March 1987.

[25] Giles Hooker and Cliff Hooker. Machine Learning and the Future of Realism. University of Toronto Press, February 2018.

[26] Paul Humphreys. The philosophical novelty of computer simulation methods. *Synthese*, 169(3):615–626, 2009.

[27] G. Indiveri. A current-mode hysteretic winner-take-all network, with excitatory and inhibitory coupling. *Analog Integr. Circuits Signal Process.*, 28(3):279–291, Sep. 2001.

[28] Kareem Khalifa. *Understanding, Explanation, and Scientific Knowledge*. Cambridge University Press, 1 edition, October 2017.

[29] Eleanor Knox. Abstraction and its Limits: Finding Space For Novel Explanation. *Noûs*, 50(1):41–60, 2016.

[30] Peter Königs. Artificial intelligence and responsibility gaps: What is the problem? *Ethics and Information Technology*, 24(3):36, August 2022.

[31] David Lewis. Causal explanation. In David Lewis, editor, *Philosophical Papers Vol. Ii*, pages 214–240. Oxford University Press, 1986.

[32] Gabriel Lima, Nina Grgić-Hlača, Jin Keun Jeong, and Meeyoung Cha. The Conflict Between Explainable and Accountable Decision-Making Algorithms. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2103–2113, June 2022.

[33] Zachary C. Lipton. The Mythos of Model Interpretability, March 2017.

[34] Ezequiel López-Rubio and Emanuele Ratti. Data science and molecular biology: Prediction and mechanistic explanation. *Synthese*, 198(4):3131–3156, 2021.

[35] David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., New York, NY, USA, 1982.

[36] Andreas Matthias. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3):175–183, 2004.

[37] R. McClamrock. Marr's Three Levels: A Re-evaluation. *Minds and Machines*, 1(2):185–196, 1991.

[38] R. C. Miall, J. G. Keating, M. Malkmus, and W. T. Thach. Simple spike activity predicts occurrence of complex spikes in cerebellar purkinje cells. pages 13–15, 1998.

[39] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, 56(4):3005–3054, April 2023.

[40] Yael Niv and Angela Langdon. Reinforcement learning with Marr. *Current Opinion in Behavioral Sciences*, 11:67–73, October 2016.

[41] Sven Nyholm. Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci. *Science and Engineering Ethics*, 24(4):1201–1219, August 2018.

[42] Chris Olah. Mechanistic interpretability, variables, and the importance of interpretable bases, 2022. An informal note on some intuitions related to Mechanistic Interpretability. Published on Transformer Circuits.

[43] Chris Olah. Interpretability dreams, May 2023. An informal note on future goals for mechanistic interpretability. Published on Transformer Circuits.

[44] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, and Charlie Shan. Zoom in: An introduction to circuits. *Distill*, 5(3):e24, Mar 2020.

[45] Samuel M. D. Oliveira and Douglas Densmore. Hardware, software, and wetware codesign environment for synthetic biology. *BioDesign Research*, 2022:9794510, 2022.

[46] OpenAI. Gpt-4 technical report. Technical report, OpenAI, 2023. Available at https://ar5iv.org/abs/2303.08774.

[47] Andrés Páez. The Pragmatic Turn in Explainable Artificial Intelligence (XAI). *Minds and Machines*, 29(3):441–459, September 2019.

[48] Paulo Pirozelli. Sources of Understanding in Supervised Machine Learning Models. *Philosophy & Technology*, 35(2):23, June 2022.

[49] Cynthia Rudin. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, September 2019.

[50] Filippo Santoni De Sio and Giulio Mecacci. Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy & Technology*, 34(4):1057–1084, December 2021.

[51] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage?, 2023.

[52] Alexander Serb and Themistoklis Prodromakis. A system of different layers of abstraction for artificial intelligence. 2019.

[53] Oron Shagrir. Marr on Computational-Level Theories. *Philosophy of Science*, 77(4):477–500, October 2010.

[54] Robert Sparrow. Killer robots. *Journal of Applied Philosophy*, 24(1):62–77, 2007.

[55] Emily Sullivan. Understanding from Machine Learning Models. *The British Journal for the Philosophy of Science*, 73(1):109–133, March 2022.

[56] Daniel W. Tigard. There Is No Techno-Responsibility Gap. *Philosophy & Technology*, 34(3):589–607, September 2021.

[57] Raymond Turner. Computational Artifacts. In Raymond Turner, editor, *Computational Artifacts: Towards a Philosophy of Computer Science*, pages 25–29. Springer Berlin Heidelberg, Berlin, Heidelberg, 2018.

[58] Robert Pershing Wadlow. Tallest man, 2010. Last measured height: 2.72 m (8 ft 11 in), on July 15, 1940, in Alton, Illinois, USA.

[59] David S. Watson and Luciano Floridi. The explanation game: A formal framework for interpretable machine learning. *Synthese*, 198(10):9211–9242, October 2021.

[60] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022.

[61] James F. Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, New York, 2003.

[62] Carlos Zednik. Mechanisms in cognitive science. In Stuart Glennan and Phyllis Illari, editors, *The Routledge Handbook of Mechanisms and Mechanical Philosophy*, pages 389–400. Routledge, London, 2017.

[63] Carlos Zednik. Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology*, 34(2):265–288, June 2021.