4.11 Interpretability

Authors: John Hewitt, Armin W. Thomas*, Pratyusha Kalluri, Rodrigo Castellon, Christopher D. Manning*

Compared to most other machine learning models, foundation models are characterized by a vast increase in training data and complexity and the emergence of unforeseen capabilities: foundation models are able to do unforeseen tasks and do these tasks in unforeseen ways. The increasing adoption of foundation models thereby creates growing desires, demands, and unprecedented challenges for understanding their behavior.

In contrast to task-specific models, foundation models are trained across vast and usually highly disparate datasets, potentially spanning many domains and modalities (see §4.2: TRAINING). Through this training, foundation models learn an exceptionally wide range of behaviors, which can vary profoundly between tasks and domains, as demonstrated by their ability to be adapted to different types of downstream tasks and to exhibit behaviors that are specific for each of these tasks (see §4.3: ADAPTATION). Take GPT-3 as an example, which was trained as one huge model to simply predict the next word in a text. While this is a very specific and simple-to-define learning task, it has enabled GPT-3 to gain capabilities that far exceed those that one would associate with next word prediction, by combining it with a vast training dataset that comprises all kinds of internet text. As a result, GPT-3 can now adapt behaviors that are clearly outside of the scope of its original training task, such as simple arithmetic and computer programming, when provided with a few training samples. This demonstrates that it is challenging to answer even the seemingly simplest question about a foundation model: what capabilities does it have?

Moreover, it is an open question to what extent these diverse capabilities rely on distinct or shared *model mechanisms*, akin to algorithmic building blocks within the model. On the one hand, foundation models can be interpreted as single models, which utilize some set of generalizable model mechanisms to perform well across tasks and domains. In this case, a full understanding of their behavior can be gained by identifying and characterising these mechanisms. On the other hand, the ability of foundation models to adapt profoundly distinct behaviors for different tasks suggests that they can also be understood as a large collection of independent expert models, each tailored to a specific task. For example, it seems unlikely that the model parameters that GPT-3 uses to do arithmetic could have much to do with the parameters used to translate from English to French. In this case, explanations of model behavior in one task are therefore not necessarily informative about behavior in other tasks. We refer to this as the *one model–many model* nature of foundation models (see Figure 23) and argue that understanding where foundation models lie on this spectrum between one and many models will be central to understanding their behavior.

Toward systematizing this area of study, we present and discuss three levels of understanding foundation models [inspired by Marr 1982]: we first discuss the challenges and opportunities in understanding *what* a model is capable of doing, then *why* it outputs certain behaviors, and lastly *how* it does it. Specifically, questions of *what* aim to characterize the kinds of behaviors that a model can perform without peeking inside the model, while questions of *why* aim to provide explanations of the model's behaviors in terms of potential causes in the data, and questions of *how* aim to understand the internal model representations and mechanisms that produce these behaviors. After presenting all three levels, we conclude by discussing potential consequences resulting from the non-interpretability and interpretability of foundation models.

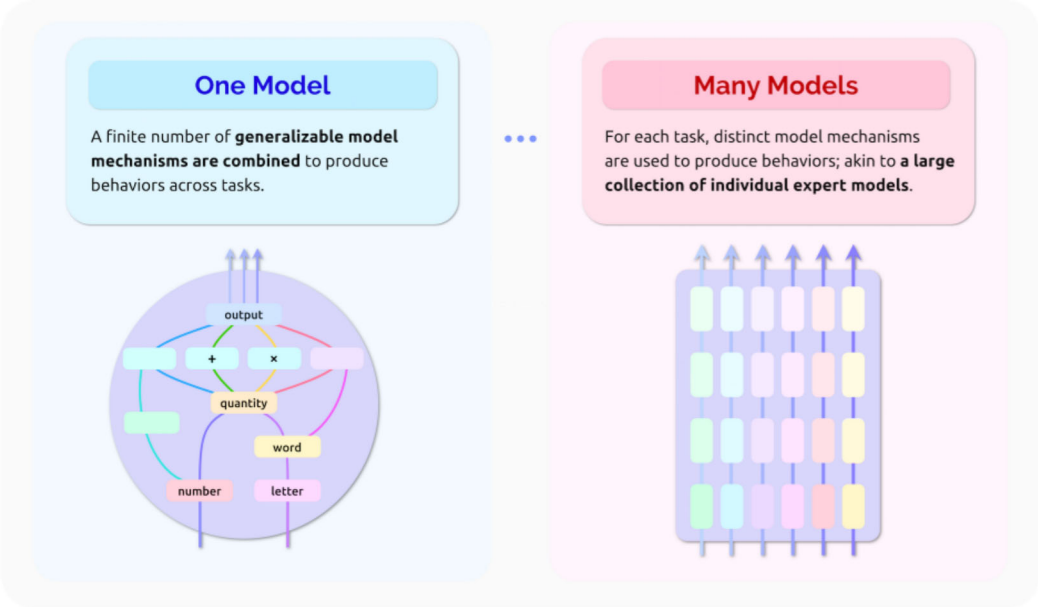

Fig. 23. The one model–many model nature of foundation models: A central interpretability question is to understand where a foundation model lies on the spectrum between *one model* and *many models*. As one model, behavior can be made interpretable by identifying and characterising the finite number of generalizable model mechanisms used to produce behaviors across tasks (e.g., mechanisms that assign meaning to words, compare quantities, and perform arithmetic). As many models, explanations of model behavior in one task are not necessarily informative about behavior in other tasks, thus requiring the independent study of behavior in each task.

4.11.1 Characterizing behavior.

The simplest understanding of a technology is widely taken to be knowing *what* the technology does. This seemingly straightforward question is significantly challenging for foundation models, due to the myriad unforeseen behaviors and tasks that these models are capable of performing.

Task-specific neural network models are trained to perform a single task in a single domain, e.g., image classification. Their task and the input and output domains are therefore clear; yet even for these models it can be challenging to know exactly what the model will do, given a particular input. For instance, model behaviors can unexpectedly differ greatly for two perceptually similar inputs [Garg and Ramakrishnan 2020; Jin et al. 2020] or two subpopulations of the same data (stratified, for example, by race or gender [Hovy and Søgaard 2015; Blodgett et al. 2016; Tatman 2017; Buolamwini and Gebru 2018]).

This challenge of characterizing a model's behavior is amplified manyfold for foundation models. The space of tasks that the model is able to perform is generally large and unknown, the input and output domains are often high-dimensional and vast (e.g., language or vision), and the models are less restricted to domain-specific behaviors or failure modes. Consider, for example, the surprising ability of GPT-3 to be trained on large language corpora and to subsequently develop the ability to generate mostly-functional snippets of computer programs. A key challenge for characterizing the behavior of foundation models is therefore to identify the capabilities that it has. Even further, for each task that a foundation model can perform, and there may be many or infinitely many, all

the challenges remain that one faces when trying to understand the behavior of much simpler, task-specific models.

Characterizing each 'task' that a foundation model can perform is further complicated by their one model–many models nature (see Figure 23). Again taking GPT-3 as an example, it was shown that it can be tailored to many tasks through simple prompting (see §4.3: ADAPTATION). Yet, each task can be specified through many possible prompts and slight variations in prompts can result in meaningful changes of model behavior. For instance, the task of sentiment classification of a movie review can be specified by presenting the movie review followed by 'Her sentiment towards the film was...' or 'My overall feeling was that the movie was...'; despite these prompts appearing to pose closely related tasks, GPT-3 will exhibit different response accuracies for each prompt [Zhao et al. 2021]. Observations like these raise important questions regarding the relationship between the characteristics of prompts and the resulting model behaviors. Specifically, can meaningfully different responses to seemingly similar prompts actually be considered as resulting from the same model or do they result from highly distinct model mechanisms, and does characterizing the behaviors of the foundation model (or its adapted derivatives) in one task truly aid in characterizing the behaviors of other possible adaptations of the model?

To identify the capabilities that a foundation model has and those it is missing, researchers can utilize controlled evaluations. Here, domain experts design prompts that are known to require a particular competence and then study the ability of a model to respond correctly to these prompts [Papadimitriou and Jurafsky 2020; Lu et al. 2021a; Kataoka et al. 2020; Wu et al. 2021c; Xie et al. 2021a; Koh et al. 2021]. For example, psycholinguists have designed prompts that require a language model to choose between a grammatically correct sentence and the same sentence with a specific grammatical inaccuracy; knowing whether the model consistently prefers the grammatically correct sentence over its grammatically incorrect counterpart tells us whether the model has the particular grammatical competence required to identify this inaccuracy [Linzen et al. 2016].

Given the huge range of possible capabilities of foundation models, and our current lack of any general method for determining a priori whether a foundation model will have a given capability, bespoke evaluations like these are crucial. They allow exploring the range of behaviors that foundation models are capable of, while requiring minimal model access: we only need to present inputs and receive model outputs, and we need not depend on access to the implementation or parameters of a model. Given the infinitely many desirable and undesirable tasks, subtasks, and behaviors that foundation models may be capable of (or incapable of), characterizing model behaviors and capabilities will be increasingly challenging and important. We believe that instead of relying on a few experts to formulate and test for possible behaviors, it will be critical to extend these types of analyses to test for many more behaviors, in part by opening up this line of exploration to diverse communities and experts in many disciplines, as well as by increasing access to and scale of these evaluations.

4.11.2 Explaining behavior.

In addition to characterizing what a foundation model is doing, one can try to characterize *why* it performs certain behaviors by providing explanations of these behaviors in terms of potential causes in the data. While current explanation approaches, which provide such explanations of behavior, can reveal qualities of inputs that affect a model's responses, they often require full access to the model to do so and are generally limited in their ability to elucidate any general model mechanisms, which foundation models use to respond to many inputs, tasks, and domains.

Current explanatory approaches can generally be understood as distinct models, which are designed to provide an explanation of particular behaviors of another *black box* model. Importantly,

these approaches are separate from the model whose behavior is analyzed, which by itself is not interpretable. This separation can be problematic, as the provided explanations can lack faithfulness [Jacovi and Goldberg 2020], by being unreliable and misleading about the causes of a behavior [cf. Rudin 2019]. Even further, unsound explanations can entice humans into trusting unsound models more than they otherwise would (for a detailed discussion of trust in artificial intelligence, see Jacovi et al. [2021]). These types of concerns grow as we transition from task-specific models towards the wide adoption of foundation models, as their behavior is vastly more complex.

Current explanatory approaches can largely be divided into either providing *local* or *global* explanations of model behavior [Doshi-Velez and Kim 2017]. Local explanations seek to explain a model's response to a specific input, e.g., by attributing a relevance to each input feature for the behavior or by identifying the training samples most relevant for the behavior [Simonyan et al. 2013; Bach et al. 2015; Sundararajan et al. 2017; Shrikumar et al. 2017; Springenberg et al. 2014; Zeiler and Fergus 2014; Lundberg and Lee 2017; Zintgraf et al. 2017; Fong and Vedaldi 2017; Koh and Liang 2017]. Global explanations, in contrast, are not tied to a specific input and instead aim to uncover qualities of the data at large that affect model behaviors, e.g., by synthesizing the input that the model associates most strongly with a behavior [Simonyan et al. 2013; Nguyen et al. 2016].

Local and global explanations have provided useful insights into the behavior of task-specific models [e.g., Li et al. 2015; Wang et al. 2015b; Lapuschkin et al. 2019; Thomas et al. 2019; Poplin et al. 2018]. Here, the resulting explanations are often taken to be a heuristic of the model mechanisms that gave rise to a behavior; for example, seeing that an explanation attributes high importance to horizontal lines when the model reads a handwritten digit '7' easily creates the impression that horizontal lines are a generally important feature that the model uses to identify all sevens or perhaps to distinguish all digits.

Given the one model–many models nature of foundation models, however, we should be careful not to jump from specific explanations of a behavior to general assumptions about the model's behavior. While current explanatory approaches may shed light on specific behaviors, for example, by identifying aspects of the data that strongly effected these behaviors, the resulting explanations do not necessarily provide insights into the model's behaviors for other (even seemingly similar) inputs, let alone other tasks and domains.

Another approach could be to sidestep these types of post-hoc explanations altogether by leveraging the generative abilities of foundation models in the form of self-explanations [cf. Elton 2020; Chen et al. 2018], that is, by training these models to generate not only the response to an input, but to jointly generate a human-understandable explanation of that response. While it is unclear whether this approach will be fruitful in the future, there are reasons to be skeptical: language models, and now foundation models, are exceptional at producing fluent, seemingly plausible content without any grounding in truth. Simple self-generated "explanations" could follow suit. It is thus important to be discerning of the difference between the ability of a model to create plausible-sounding explanations and providing true insights into its behavior.

4.11.3 Characterizing model mechanisms.

Deep understanding of systems is generally taken to mean understanding *how* a system performs: which knowledge and mechanisms does it contain, and how are these assembled to form the whole?

If this is indeed possible, characterizing the representations within foundation models and the mechanisms that operate on them will be central to satisfying the desire to thoroughly understand these proliferating models; and whether these mechanisms are many and specific or few and generalizable, they are at the core of the ability of foundation models to adopt a wide range of behaviors in varied tasks and domains.

To make the notions of model representations and mechanisms concrete, consider a simple behavior exhibited by GPT-3: It was quickly observed *what* GPT-3 did when provided with examples of the addition of small numbers and then queried to perform addition of two new numbers: with high probability, it predicted the correct result of the addition [Branwen 2020; Brockman 2020]. When asking *why* GPT-3 performed as it did, one could find evidence in the input, like aspects of its prompt that highly affected its response (these might be the two numbers to be added, though not necessarily), or aspects of GPT-3's training data that affected its response (these might be examples of addition, though not necessarily). Delving into the model, we may envision a deeper understanding of the mechanisms that GPT-3 uses to add a specific pair of numbers and the mechanism that it uses to add other arbitrary pairs of numbers. We may also envision a deeper understanding of whether these mechanisms are similar to the mathematical notion of 'addition' or merely correlated with this notion.

By understanding individual model mechanisms, we can build up a compositional understanding of complex behaviors of a foundation model. A task slightly more complex than the addition of numbers is solving mathematical word problems, in which numbers come with units and the problem is presented in natural language. Once we understand the mechanism (or mechanisms) by which a model performs addition, we can investigate whether this mechanism is used as an intermediate step in solving word problems. If the addition mechanism is used, we have built up our understanding of how the model solves word problems, we have increased confidence that the foundation model generalizes the notions of quantities and addition (not another correlation or heuristic), and, furthermore, we have increased confidence in our ability to predict the model's *why* (which parts of the inputs it is attending to) and the output's *what* (addition of two numbers). If the addition mechanism is not used, we may retain a healthy skepticism that this is truly addition, and we can investigate which representations and mechanisms are used instead.

It is important to be aware that there are many potential cases of more complex and concerning model mechanisms, for instance, the estimation of race from the characters in a name, or the pixels in an image. Establishing evidence of such a mechanism in a foundation model and its use can support a moral or legal responsibility to ban the model from tasks like predictive policing, marketing, loan applications, and surveillance at large.

A plethora of methods have emerged to investigate these internal aspects of neural network models. Typically, these approaches separate the model into nodes (e.g., neurons, layers, or parts of layers), then interrogate either the representations captured in nodes or the mechanisms by which nodes are assembled. Some approaches are hypothesis driven: by hypothesizing that nodes may capture certain information (e.g., a grammatical feature of a word, or the race of a person), one can probe all nodes to quantify how much of that information they make available [Alain and Bengio 2016; Veldhoen et al. 2016; Belinkov et al. 2017; Adi et al. 2017; Conneau et al. 2018; Hewitt and Liang 2019; Hewitt and Manning 2019; Voita and Titov 2020; Pimentel et al. 2020]. Other approaches build on explanatory methods, and, instead of identifying which data cause a certain behavior, they seek to identify which data cause a certain node to activate, or which nodes cause another node later in the model to activate, thereby uncovering collections of model representations and mechanisms [Olah et al. 2020; Mu and Andreas 2020; Carter et al. 2019; Goh et al. 2021]. Taken together, these approaches inspect the interior of models and provide a basis for the ongoing explorations of the behavior of foundation models. Yet, the number of potential representations and mechanisms within foundation models is vast, particularly given their one model–many models nature, and these types of approaches often only capture a small slice of a model's interiority. It is thus an open challenge to expand the discovery of representations and mechanisms and to elucidate those that are most relevant or general for model behavior. As with many approaches to interpreting foundation models, these types of explorations will benefit from

including and supporting more diverse and interdisciplinary investigators and from more accessible, flexible, and scalable methods of discovery.

In summary, we believe that the one model–many models nature of foundation models (recall Figure 23) provides novel opportunities and challenges for current interpretability research: there are many adaptations of a single foundation model, and we simply do not know the extent to which they share common mechanisms. To the extent that mechanisms are shared, understanding foundation models may be a tractable problem of characterizing these mechanisms and their relations. To the extent that mechanisms are independent, each adaptation of a foundation model must be analyzed independently, leading to profound uncertainty about the nature of any new adaptation of the foundation model.

4.11.4 Impacts of non-interpretability and interpretability.

Lastly, we would like to highlight that the wide adoption of foundation models is at odds with a recent plea of many interdisciplinary researchers not to use complex black box models for high stakes decisions [e.g., Rudin 2019], but instead to focus on the long-standing development and application of more intrinsically interpretable models.

In the midst of these pleas, work aimed at interpreting foundation models is a double-edged sword. Large machine learning models, and now foundation models, are most often deployed by powerful corporations and institutions, and incremental advances in interpretability can be exaggerated to 'ethics-wash' and continue use of models as though they have *achieved* interpretability, belying the reality that they remain far below traditional standards of algorithmic interpretability. Moreover, when approaches to interpretability regularly presume easy access to models and their implementation and parameters, interpretability can serve not only as cover for powerful institutions but also centralize model knowledge in the same hands. For those working toward the interpretability of foundation models, it is a responsibility to consistently ask whether one is working toward making foundation models *interpretable to researchers and model owners* or *interpretable to everyone*.

Simultaneously, to the extent that foundation models are already being deployed, work on interpretability presents unique opportunities to shift knowledge of foundation models, and thus power, back to datafied and evaluated peoples. Interpretation can facilitate the discovery of societally salient aspects of models. More radically, work creating accessible methods that allow anyone to interpret the behavior of foundation models shifts power to diverse peoples, creating opportunities to investigate models, opportunities to discover aspects of models important to individuals or their communities, and opportunities to meaningfully consent to, improve, or altogether contest the use of foundation models. Finally, it is important for researchers to view the interpretability of foundation models as not only a goal, but a question: research can explore and assess whether the lack of foundation model interpretability is intrinsic and should be deeply studied and widely known as a serious issue discouraging use (or increasing regulation) of these systems, or whether it is possible for future foundation models to uphold a high standard of interpretability for all.