2.5. Governance via Explainability

David Danks
University of California, San Diego

1. Introduction

Successful governance of AI systems requires some knowledge of how, and more importantly *why*, the system functions as it does. If I do not understand why a loan approval algorithm denied me a loan, then I cannot change things in hopes of a better outcome next time. If regulators do not understand why a self-driving car makes certain decisions, then they cannot safely determine when the car should or should not be used. If a doctor does not understand why a medical AI made a mistaken diagnosis, then she may not be able to provide feedback to improve its future performance. And similar observations could be made across a range of different domains and AI systems. In general, if relevant stakeholders—developers, users, citizens, and so forth—fail to understand why the system works as it does, then many aspects of governance either will be infeasible or will require costly trial-and-error.

This deeper understanding is particularly important in the context of AI since prior testing might not be feasible. One of the key motivations for deploying an AI system is that it is hopefully intelligent enough to adapt to novel, unexpected, or unforeseeable circumstances. That is, we want to use AI systems *precisely* when we cannot measure performance in all relevant contexts prior to deployment. Standard approaches to governance using mere reliability will almost always be insufficient for AI systems since they will inevitably be used in unexpected situations. Instead, we have even more reason to require an understanding of the functioning of the AI system—said differently, an understanding of why it works as it does.

This kind of understanding is, in everyday life, usually provided by explanations. I could ask a loan officer to explain my denial; regulators could ask human drivers why they made that choice; or a doctor could explain her reasoning to a review board. So if we are interested in AI governance, then we might naturally be interested in AI systems that are, in some sense, suitable for explanations. Moreover, there has been significant research on these explainable AI (XAI) systems over several decades; the underlying technologies are starting to mature. But while this connection between XAI and AI governance is intuitively appealing, matters are not so simple, precisely because explanations, XAI, and governance are all more complex than these initial observations suggest.

This chapter aims to show how different kinds of explainability can be used to support different functions of AI governance. While there is heterogeneity in explanations and XAI, there is a universal feature of all explanations that can be used to provide concrete guidance about how XAI can, and sometimes cannot, support AI governance. In particular, the quality of explanations depends on whether they support the goals of the explanation recipients. The result of this analysis is a concrete framework for "AI governance via explainability" or "explainability for AI governance," rather than specific policy or process recommendations. This framework is complex; there is no simple way to govern AI via explainability given its many different uses, contexts, and stakeholders. However, this complexity need not lead to paralysis; the motivating intuition is correct that explanations can sometimes improve AI governance.

This chapter has much less discussion on the topic of "explainability *through* governance" or "explainability as a goal of governance." There are many different methods and frameworks that have been proposed to help guide the development of XAI systems, some of which presumably

count as forms of governance to reach the goal of XAI. At the same time, though, explainability is very rarely an end in itself; rather, explainability is a goal because it could help to increase use, performance, trust, control, or some other more fundamental goals that are central to successful (from a societal perspective) governance. Explainability through governance is important primarily because it can enable governance via explainability, and so the focus of this chapter will be on the latter possibility.

Section 2 explores the intuitive connection in more detail to show why explanations seem well-suited to support or enable governance, particularly for AI systems, but also why the details matter. Sections 3 and 4 then respectively consider XAI and explanations in more detail. Section 5 returns to the intuitive connection to develop the framework for AI governance via explainability. While AI governance cannot be guaranteed simply by using explainable AI systems, an appropriate use of explanations can significantly advance the governability of AI. Section 5 focuses on high-level considerations about governance without being tied to any particular proposed or current law, regulation, or policy. This agnosticism ensures that the framework can be used both to evaluate current ideas, and also to guide future ones. Section 6 provides some concluding thoughts. We begin, though, with a more careful consideration of the intuitive connection.

2. A *prima facie* connection

As noted above, a salient feature of AI systems (in contrast with non-autonomous systems) is the difficulty of knowing or predicting how they will behave in novel situations or contexts. If the desired system performance could be fully specified ahead of time, then there would be no reason to use anything that might be called 'AI' (rather than a much simpler algorithm or computational system), as one could simply "hard code" the desired behavior. Part of the usefulness of AI systems is precisely that they can be flexible and surprising, hopefully in positive ways. For example, successful self-driving cars must do more than follow simple patterns, but rather adapt to the constantly changing roadways. And since AI systems will be used in novel or surprising situations, "mere" reliability information is insufficient for appropriate use and governance. The insufficiency of classical reliability information is exacerbated when, as frequently occurs, people have a very different understanding from the AI of which situations are novel.[1]

Instead, we need to know *why* the system behaves in particular ways. Explanations are often proposed to be answers to so-called why-questions, such as "why did she choose that career?" or "why did the bridge support fail?"[2] Hence, we might naturally look towards explainable AI (XAI) as more easily governable, or perhaps even the only kind of governable AI.

The field of explainable AI dates back several decades, and has experienced a renaissance in recent years. There are multiple kinds of AI that have been described as "explainable"; XAI is not one single technology. For example, a loan approval algorithm could be XAI if (a) it self-generates explanations of approve/reject decisions; (b) data scientists can analyze it to understand why it made particular approve/reject decisions; or (c) loan applicants can interpret it in ways that (seemingly) yield why-information. Of course, the same loan approval algorithm could fit multiple of these characterizations. XAI will be the focus of the next section; for now, we only need the observation that there are many kinds of XAI.

Given this diversity, AI governance via explainability will depend on a better understanding of the nature of explanation. Explanations do not merely *describe* some event or phenomenon, but rather provide (in some sense) an account of *why* it occurred. A theory of explanation must

explicate what more is required for something to be an explanation, rather than a mere description. This explication could be either philosophical or psychological, depending on whether the focus is, respectively, what an explanation ought to be (normatively, rationally), or what an explanation actually is for humans (descriptively, cognitively).

Section 4 will consider theories of explanation in detail, but a high-level overview will be useful in the meantime. One classic philosophical theory of explanation is that explanations are simply predictions where the outcome is known;[3] that is, an explanation of some event is a description of the conditions from which one can predict that event (and we know that it actually occurred). In the context of AI, this idea would suggest that an XAI system would only have to provide the (relevant) input data that produced the output. This natural idea will not work without significant amendment, however, since description of the inputs provides only the initial conditions for the event, not the explanation of it. For example, a list of input symptoms will not provide an explanation for why a medical AI system diagnosed someone with a disease. More generally, a theory of explanations must account for the fact that some information conveys *why* something happened (i.e., is a genuine explanation) while other information only conveys *that* something happened or the conditions for it to happen.

Matters become even more complex with psychological theories of how people actually use explanations, or decide whether a description counts as an explanation. In particular, explanations might appear to be backwards-directed, in the sense that they give a retrospective account of why something happened. Explanations seem, on the surface, to be about the past. However, it turns out that they have a significant forward-directed psychological function: explanations can help the recipient better predict and respond to similar situations in the future.[4] Psychologically, explanations are not just about what did happen in the past, but also about what might happen in the future. Whether considered from a philosophical or psychological perspective, explanations are complex objects.

This section has outlined a *prima facie* plausible argument: AI governance requires understanding "why?"; explanations answer why-questions; therefore, AI governance requires explainable AI. At the same time, even a high-level exploration of the steps in this argument has revealed complexities and nuances that are obscured by the *prima facie* formulation, and many more issues lurk just beneath the surface. We must dig deeper than this overly-quick two-premise argument. In particular, each different type of XAI requires a substantive theory of explanation in order to be usable. The features that are shared by different theories of explanation thereby determine what features must hold of any XAI system. That connection prompts the exploration of both philosophical and psychological theories of explanations to find those common elements.

3. A taxonomy of explainable AI (XAI)

The idea that an AI system might, or should, be explainable (in some sense) has a long history, dating back at least to the 1970s.[5] Many of those early examples were expert systems that were supposed to assist human decision-making, or perhaps replace human decision-making only after validation (partly) on the basis of expert knowledge. The desire for explanations was thus largely driven by skeptical humans who questioned the possibility that an AI system could help or replace them. That is, XAI was needed to convince humans that the AI system "knew what it was doing." XAI faded as a central topic with the rise of "big data" machine learning systems, including (but not limited to) advances such as deep neural networks. This rise changed the justification for AI systems to their ability to identify patterns in data that were unnoticed or

unlearnable by humans. And since part of the appeal of AI systems was exactly their ability to understand, predict, or control the world in ways that (seemingly) exceeded human powers, explainability no longer seemed so important. However, many recent events have highlighted the costs of this shift. For example, when AI systems understand the world in different ways than people, then it can be quite difficult to accurately predict or determine the contexts in which the AIs will fail. These non-explainable systems also typically do not provide useful insights that can be applied elsewhere; people instead must simply accept (or not) the AI system output. For these reasons (and many others), XAI has reemerged as a major topic of research in recent years.

The arguments in this chapter largely do not depend on technical details about XAI systems, so I will not provide a technical survey of the many different XAI methods. Interested readers should consult one of the many such surveys that are now available.[6] At the same time, I will use particular XAI approaches or techniques as examples (without much detail) in order to show how the topics in this chapter connect with that technical literature.

Any overview of XAI is complicated by the relative lack of agreement about terminology. There is a set of interconnected concepts and terms—explainability, intelligibility, interpretability, transparency, understanding—that are not used consistently across the field. For example, one person's "explainable AI" could be another's "interpretable AI." This section will not attempt to adjudicate those terminological disputes, but will rather focus on the different functional types that fall under the broad label of 'XAI'. As a result, some instances of XAI (according to this chapter) might be called something different in other contexts. This terminological agnosticism will enable progress on the relationship between AI explanations and AI governance, and might even provide a new way to draw principled distinctions between different types of XAI.

With these caveats in mind, one can distinguish at a high level between three different, not mutually exclusive, types of XAI. These three types can blur together in some cases, but they provide a useful taxonomy for thinking about AI governance. The first type of XAI—what can be called *explanation-generating AI*—is one that is itself capable of providing an explanation of its behavior when queried or probed. For example, a loan approval algorithm that recommends approval for an applicant might provide an accompanying explanation for its judgment, such as the key counterfactuals about what changes would have led to rejection of the application.[7] These system-generated explanations could potentially be generated by a sub-system that analyzes the original AI; this kind of add-on module for explanation can enable one to convert many different AI systems into XAI ones, potentially even in a post hoc manner.[8] The key characteristic for explanation-generating AI is that the system itself produces the explanation. The humans who develop, use, or otherwise interact with the AI need not do any particular cognitive work.

One challenge in assessing explanation-generating AI is that they are often thought to provide *justifications* for particular judgments, not merely explanations for them. For example, a loan approval algorithm of this type is sometimes expected to explain not only why it provided a particular judgment, but also why that judgment is legally, morally, or socially acceptable. There are three main reasons why this chapter will largely set aside the potential justification-generating capabilities of some XAIs, and focus purely on their explanation-generating capabilities and resulting implications for AI governance. First, explanations and justifications are simply two different goals for an XAI system. In general, explanations purport to tell us why something happened (or did not happen) while justifications purport to say why something was right (or permissible or acceptable). That is, explanations are largely descriptive, while

justifications involve a significant normative component that defends the action as appropriate. Second, there is often significant disagreement about which normative standard should be used, and so disagreement about how to evaluate an AI-generated justification. Third, the other two types of XAI systems are used much less frequently to try to produce justifications for system outputs, and so a focus on justifications would omit a large number of XAI systems.

Returning to the three types of XAI, the second type—*human-explainable AI*—arises when an appropriately knowledgeable or trained human is able to generate explanations, usually for themselves, of the AI behavior. Many canonical examples of XAI fall into this type. For example, shallow (i.e., few-layer) decision trees are widely thought to be explainable systems, but explanations of their output or behavior (e.g., "this person was approved for a loan because their credit score was high and their debt-to-income ratio was low") are actually generated by people reflecting on the model, not the AI systems themselves. A decision tree does not itself provide an explanation; the human plays an integral role in the production of an explanation for a decision. Other popular XAI techniques of this type aim to extract low-dimensional approximations of the actual high-dimensional model.[9] Again, the AI system does not itself generate any explanations, but rather provides information that is useful for a knowledgeable human who is trying to make sense of the system performance. Human-explainable AI is clearly dependent on the knowledge and skills of the relevant human. Most work has assumed that the relevant human is generally knowledgeable about algorithms and computational models, and so can understand things like low-dimensional approximations without further training. At the same time, research on human-explainable AI usually requires that the required training is not *too* specialized, so some AI systems (e.g., deep neural networks) are consistently classified as not human-explainable even if a few people actually could generate an explanation from them.

Finally, a third type of XAI—*human-interpretable AI*—is one that exhibits patterns of behavior for which untrained people can generate satisfactory stories. These stories can "explain" the AI in the sense of capturing the patterns of AI behavior, while not necessarily counting as "explanations" on standard philosophical or psychological theories of explanation. Research on human-interpretable AI focuses on shaping or constraining the AI's behavior so that humans without specialized knowledge can produce "as if" stories, perhaps in the same ways that people generate stories about one another to explain behavior. For example, a robotic system is sometimes described as XAI if humans can understand the robot's behavior "as if" it had beliefs, desires, and other mental states.[10] Of course, those "explanations" might be entirely wrong about the actual inner workings of the robot; there might be no representations or content in the robot that fit our naïve understandings of beliefs, desires, and so forth. Nonetheless, if (untrained) people are able to generate stories that "make sense" of the AI behavior, then there is a sense in which the AI is explainable. At the least, these stories might enable people to achieve many of the goals that explanations normally support.

Each of these types of XAI could be useful in a particular context, and could improve or increase AI governance. One significant challenge, however, is that *each of these types of XAI requires a substantive theory of explanation*; the discussion in this section has taken for granted that we know what should or does count as an explanation. For example, consider explanation-generating AI: without a substantive theory of explanation, the developer would not know what kinds of explanations should be generated by the AI system, nor how to evaluate whether the AI system actually succeeds in achieving (this type of) explainability. Similar observations arise for the other two types of XAI, and so there is seemingly an explosion of types of XAI: the three here, multiplied by all of the different substantive theories of explanation. Such a proliferation of

types poses a significant barrier to AI governance through AI explanation, as there may be simply too many different permutations. Successful governance requires a response to this challenge, but that response will require closer examination of the nature of explanations.

4. What is an explanation?

Both philosophical and psychological theories of explanation are relevant for potential AI governance. The former type of theory aims to articulate what an explanation ought to be, in some normative sense; what ought to be the features or characteristics of something that truly does answer a why-question? The latter type of theory characterizes how (purported) explanations actually function in human cognition; how do people use things that seem to be explanations in understanding, reasoning about, and acting in the world? And how do people determine that something might be an explanation? The philosophical and psychological theories can clearly diverge from one another: people might not correctly identify and use "real" (according to some philosophical theory) explanations, and a philosophical theory might not make any reference to an explanation's cognitive impacts. Nonetheless, we should plausibly expect *some* connections between philosophical and psychological theories of explanation, much as we expect connections between such theories of causation, action, agency, and so forth. More importantly, both types of theories are relevant for governance via explainability, as both the information in an explanation (normative aspects) and people's responses to explanations (descriptive aspects) will impact AI governance.

4.1 Philosophical theories of explanation

At a high level, philosophical theories of explanation divide into two types—realist and pragmatic—depending on whether the quality of the proposed explanation is evaluated based on its accuracy or truthfulness, or instead based on its pragmatic value to the recipient of the proposed explanation. That is, these two kinds of theories differ based on whether explanations should mirror (in some sense) reality in the right ways, or whether they should support people's cognitive needs in the right ways. Many explanations will satisfy both requirements (truthfulness *and* helpfulness), but some explanations, including some kinds of XAI, might satisfy only one.

Realist philosophical theories of explanation hold (roughly) that an ideal explanation will articulate all-and-only the actual, true reasons why the explanandum—that is, the thing to be explained—occurred, perhaps in the particular way that it did. For these kinds of philosophical theories, a proposed explanation that gets the facts wrong is simply not an actual explanation, regardless of whatever other benefits might result from someone receiving it. For example, an explanation of why a tree's leaves are green should make reference to chlorophyll absorbing red and blue parts of the visible spectrum. In contrast, the proposal that magical fairies paint the leaves green when no one is watching would equally well enable correct prediction, generalization, and so forth, but would not be an explanation.

Of course, this simplistic characterization in terms of true facts cannot be the full story. In particular, a realist philosophical theory must provide the restrictions on a set of facts (about events, laws, causal relations, and so forth) that must hold for it to actually answer a why-question. For example, some accounts might require that an explanation include (necessary) laws of nature[11] or causal relations and structures,[12] or that it provide a unification of multiple events or phenomena,[13] or some other additional criteria beyond simply providing relevant facts about the events leading up to the explanandum. Moreover, explanations can sometimes seemingly include false-but-approximately-correct claims, as when one explains the changing tides by

appeal to Newtonian gravitational forces (plus the changing location of the moon and other facts), rather than the laws of general relativity. An explanation thus must be the right (in a sense to be considered shortly) set of true facts, not just any arbitrary collection of true facts.

In contrast with realist theories of explanation, pragmatic accounts focus on the functional role that explanations ought to play for the recipient.[14] In general, explanations enable people to better understand what occurred, and pragmatic theories hold that this impact is the core characteristic of an explanation. That is, explanations are whatever increases understanding, even if it fails to mirrors reality. As a result, pragmatic theories allow for the possibility that false statements can nonetheless explain[15] (e.g., if they provide a useful analogy, or a useful "as if" story as for human-interpretable AI). Moreover, understanding critically depends on the goals and/or context of the recipient, and so pragmatic theories of explanation argue that there is a component to every explanation that necessarily depends on features of the recipient. On a pragmatic account, the question "is this series of statements an explanation?" simply cannot be answered without knowing about the goals and/or context in which those statements are provided.

One obvious implication of a pragmatic theory of explanation is that the exact same statements could be an explanation for one individual but not for another.[16] For example, an explanation in terms of quantum mechanics might be useful for a physicist but not a young child; more relevantly here, an explanation in terms of a complex machine learning algorithm might be useful for an AI researcher but not a member of the general public. This audience-dependence is also endorsed by proponents of realist accounts of explanation, though primarily because of the pragmatics of conversation, not any necessary aspect of explanations themselves. That is, the proponent of a realist theory can acknowledge that we give different explanations to a child and a quantum mechanic, but reject the idea that this explanatory practice thereby tells us anything interesting about the nature of explanations.

The clearest point of departure between the types of theories is whether radically false (i.e., not even approximately true) statements can be part of an explanation: realists say "no" while pragmatists say "yes, if it increases understanding." Of course, radically false statements will often not contribute to understanding, so we should probably expect that most explanation will involve (approximately) true statements. Nonetheless, the question of whether radically false statements can *ever* be part of an explanation highlights the different grounds for explanations—accuracy vs. understanding. This question is particularly salient for human-interpretable AI systems, since the stories that people generate might involve exactly these kinds of radically false statements (e.g., "the robotic car *believes* that there are people in the road"). This question is also salient when people look to explanation-generating AI systems for justifications, since justifications are rarely evaluated based on their helpfulness.

As noted above, the diversity of normative theories of explanation potentially poses a challenge for AI governance, since there could be a problematic proliferation of XAIs. However, if there are features or properties that are shared by (almost) all substantive theories of explanation, then those can be used for AI governance via explainability, regardless of the particular type of XAI. One can remain agnostic about which normative account is right, and instead simply use the shared features and properties. The resulting methods and practices would have force and legitimacy across a wide range of settings, commitments, and approaches. Agnosticism about the "true" nature of explanation can thus be seen as analogous to agnosticism about "the good life" that underlies many governance systems for political life (in value-pluralistic societies).

One might reasonably wonder whether there are *any* features that are shared by all substantive normative theories of explanation. I propose that the (broadly understood) goals of the explanation recipients are necessarily relevant for each of these types of theories. Obviously, the recipients' goals are critically important for pragmatic theories; one cannot know whether something contributes to understanding without knowing why the recipient wants to understand. In contrast, realist accounts seemingly make no explicit reference to goals, but I contend that they involve an implicit dependence on goals.

In particular, recall that realist accounts must be supplemented in some way to indicate which sets of (approximately) true statements constitute an explanation. This addition could be provision of a measure for "approximate" truth, or restriction to certain sets of causal relations, or specification of neighboring theories that are unified via the proposed explanation, or many other supplements. In each case, though, the justification of a restriction will depend on the (broadly understood) goals of the recipient. For example, a restriction to specific causal relations might be appropriate only if the goal is control (which requires causal knowledge). Or what counts as an acceptable level of approximation will depend on goal-specific features. These implicit goals could be incredibly broad such as "know more about the world," but even that goal still contrasts with other possible goals (e.g., "better control the world"). Moreover, these goals are not tied to particular levels of description;[17] this goal-dependence is *not* an instance of the previous observation that conversational pragmatics can influence what explanations we happen to offer. Rather, full specification of a realist theory of explanation requires (implicit) specification of the recipients' goals in order to ground or justify the necessary, additional constraints on (approximately) true statements. Hence, we can see goal-dependence as a shared feature of substantive philosophical theories of explanation.

4.2 Descriptive theories of explanation

Now consider descriptive theories of explanation: what role do explanations play in human cognition? If explanations are to improve AI governance, then their use should depend on how the cognition of relevant stakeholders (e.g., developers and users) is influenced by those explanations. Of course, people's cognition will change after receiving *any* set of statements; for example, if I think that some statements are true, then I will have new beliefs, additional inferences from those new beliefs, and so on. The challenge for psychological theories of explanation is to determine what additional cognitive changes result when one receives an *explanation*, not just a set of statements.

One change that is not particularly relevant is people's subjective experience of liking (or not) a putative explanation. Cognitive changes and the phenomenology of explanations could presumably separate: something could provide cognitive benefit without people liking it, and vice versa (as is frequently found in pedagogical studies, or when people like something solely because it is familiar). For the purposes of AI governance, the cognitive impacts are the most relevant, rather than the experiential ones. Explanations can presumably provide a useful mechanism of AI governance only through changes in people's subsequent decisions and reasoning, rather than through a momentary good or bad experience (though with the caveat that a sufficiently bad subjective experience might lead someone to ignore the explanation). Whether someone "likes" an explanation—or even is willing to call something an 'explanation'—is not the focus here; the question is how people think and decide differently as a result of the explanation.

One important feature of the cognitive impact of explanations is that they alter, hopefully for the better, people's *future* reasoning, prediction, and action, not only their knowledge of the past.[18] The statements in an explanation almost always refer to past features of the world, including both the past state of the world and the scientific laws and causal structures in place at the time. If one is provided this set of statements in a non-explanatory context (e.g., if the statements are a mere description), then one's cognition about the past will change since something is learned, but cognition about the future will not significantly shift. If these statements are instead presented as an *explanation*, then numerous studies have shown that one's cognition about the future will also change.[19] For example, suppose I see a fallen tree and am told "there was a beetle infestation last year." If this claim is presented as merely a description of the forest, then I simply learn about some events from last year. If that statement is instead presented as an explanation, then I infer more, such as that beetles are the kinds of things that can lead to fallen trees. Future predictions will change in light of this new knowledge in ways that go beyond the impact of the facts about last year.

The future-directed impacts of explanations can be understood in terms of generalization. Explanations indicate the features of the world that are relevant to understanding why something occurred, and so convey information about which features are likely to be relevant in future contexts. When the fallen tree is explained in terms of a beetle infestation, then if I care in the future about predicting or preventing dead trees, then I should seek information about beetle infestations. While various descriptive theories might differ about the exact impacts on future cognition, they share the conclusion that explanations are not purely backward-directed but have significant forward-looking impacts.

As this example shows, the descriptive quality of an explanation will depend on whether it enables the right kinds of future cognition. That is, whether something is a good explanation (in descriptive terms) will depend on whether it provides the information for the recipient to succeed at relevant future cognitive tasks. But the relevant future cognition will depend on the goals and needs of the recipient: something could be a good explanation for certain goals, but if I never actually encounter those corresponding cognitive tasks in the future, then it is not helpful for my particular cognition (and so not actually a good-for-me explanation). These goals could be quite broad and vague (e.g., "be prepared for surprises in the future"), but the psychological quality of an explanation nonetheless depends on them.

Goal-dependence or -sensitivity thus emerges as one universal feature, perhaps one of many, across essentially all substantive theories of explanation, whether philosophical or psychological, though the details of that dependence can vary. The next section shows how to use this universal feature to better understand how XAI might, and might not, be used to improve AI governance. If there are other universal features of substantive theories of explanation, then those could also be incorporated in similar ways.

5. Governable AI via explainability

We start by considering some of the goals of AI governance, as those will constrain the type of explainability that might be useful for governance. In particular, the goals of AI governance might require certain kinds of explanations, and thus certain kinds of XAI, at least to the extent that we care about governance via explainability.[20] I adopt a relatively general notion of governance as the mechanisms that steer the governed towards desired outcomes and targets, similar to a forward-looking version of notions such as "accountability as a practice."[21] Governance on this broad conception has the overall function of providing some level of

assurance that our AI systems will bring about the outcomes we want, and also that appropriate responses will be taken when they fail to do so. In this section, I consider the implications for XAI of four different potential requirements for this type of broad AI governance: system prediction, system control, failure signals, and proper incentives. Of course, these four form only a partial list; no claims of completeness are made or intended, though these four features will arguably be relevant for any governance process. How should these requirements, and the corresponding goals to achieve each, constrain the types of XAI that might be developed or deployed?[22]

The first requirement was mentioned at the start of this chapter: making predictions about system performance in novel circumstances. Appropriate governance mechanisms require the capability to make (noisy, defeasible) inferences about the likely AI performance in new situations so that the appropriate contexts or scopes for its use can be determined. Prediction for novel circumstances is critical to address this governance challenge. Explanations can clearly support predictions in novel circumstances, but they need to be either realist explanations or pragmatic ones with this goal. Explanation-generating and human-explainable AI systems are thus likely to be helpful. In contrast, human-interpretable XAI systems are typically built so that people can construct stories for normal operation, and those stories will not necessarily provide accurate predictions in novel contexts (whether because the stories are not accurate, or they have the wrong pragmatic goal). For example, I might interpret a robot as if it has human-like beliefs and desires, only to be quite surprised at its behavior in new situations if it does not *actually* have beliefs and desires. Regardless of the type of XAI, it should lead to explanations (or stories) that prioritize the goals of system deployers. This requirement is needed for governance over contexts of use, and deployers are the individuals who have the largest impact on that aspect of AI systems. Explanations that instead help users make predictions, for example, would not necessarily support this governance function since users have relatively little control over deployment contexts.

A second goal is making predictions given interventions or changes to the AI system, relevant contexts, or human users. Governance requires mechanisms that can shift the behavior of the governed system in appropriate ways, which presupposes some ability to estimate how the system might respond to such changes. Governance mechanisms should only prescribe various adjustments to an AI system given reasonable inferences about the results of such changes. Prediction given interventions is importantly different from prediction given observations. One can predict that the current temperature outside is cold by observing people wearing heavy jackets, but intervening to force people to wear heavy jackets in summer will not lower the temperature. Both kinds of prediction—from observations and from interventions—are important for the design and use of successful AI governance, but they must be separately supported. Similarly to the first goal, explanation-generating and human-explainable AI systems are likely to be helpful,[23] but human-interpretable AI systems will not necessarily provide appropriate explanations for this requirement, unless those stories happen to correspond to the actual causal structure of the AI system. In contrast with the first requirement, these explanations (or stories) should be appropriate for both deployers and users, as both are likely to be in a position to change or impact the AI system.

The third governance requirement is knowledge or understanding of indicators of failure or problems, as this awareness is a prerequisite for appropriate monitoring and oversight. AI systems deployed in open contexts will inevitably surprise us, whether in good or bad ways. Their full performance profile will almost never be known in advance of their use, and so

governance requires mechanisms to detect problematic AI behavior. Hence, governance via explainability should support the goal of appropriate detection capabilities, where this goal is shared by both regulators and users. One key criterion for fault detection is to distinguish between errors that should be corrected and the inevitable failures that are simply part of normal operation in a noisy world. For example, a loan approval algorithm will surely not be perfect, but its failure can have different sources. Some of its judgments will be wrong simply because they are based on imperfect, partial data, while others might be wrong because of systematic (and legally problematic) biases in the algorithm. A good governance mechanism should minimize or mitigate the latter kind of errors, but that requires the ability to distinguish between these kinds of errors. For this requirement, explanation-generating and human-explainable AI systems can provide the required information for regulators and users. More interestingly, human-interpretable AI systems can also provide useful stories, though only if those stories are tied to the identification of appropriate behavior. Human-interpretable AI systems will not necessarily enable one to know how to respond to failures, but they can help to identify those failures.[24]

The fourth requirement for AI governance extends out from the technology to include the humans involved in its design, development, deployment, and use. In particular, people will frequently be "in the loop" with AI systems, and so those people's actions must be taken into consideration when aiming for governance. Even a well-designed AI system could lead to problematic outcomes if people deliberately misuse it. For example, racist people using an unbiased loan approval algorithm could do a great deal of harm that proper governance should minimize or mitigate, but the focus should be the people not the AI. However, governance mechanisms will typically not be able to constantly monitor the people, and so successful governance requires the creation, implementation, and maintenance of proper incentives to ensure appropriate behavior. One might note here that XAI does not seem particularly relevant to proper incentives, and that observation is actually the point. This particular requirement for AI governance is included precisely because the move from AI to XAI does not advance it in any substantive way. Increased understanding of the AI system will probably not help to understand or create proper incentives. Explanations of the AI system, regardless of type or source, will simply not help one to understand how the broader social system could (or should) be changed, particularly when there are significant systemic biases in our data or society. Although governability can be improved via explainability, XAI is not a panacea for all challenges of AI governance.

6. Conclusions

As AI system proliferate in number, authority, and autonomy, there is increasing need for mechanisms to govern them in various ways. Explainable AI superficially holds the promise to enable the necessary governance; one might even be tempted to require all AI to be XAI in order to ensure that the systems are governable. This temptation is understandable, but also ultimately misguided. As demonstrated in this chapter, AI governance via explainability is a complex possibility that depends on the type of XAI, type of explanation, and relevant requirements or goals of the governance effort. At the same time, this complexity need not be overwhelming or paralyzing: there are commonalities within each of these dimensions that can enable us to provide concrete guidance, primarily shaped by the goals and needs of the recipients of the explanations.

This chapter has provided a framework for pursuing AI governance via explainability, but it clearly has not been exhaustive. For example, one reason to have a governance system is to

increase trust and utilization of a system: if one knows that there are mechanisms in place to nudge the AI towards better (in some sense) behaviors, then one is more likely to trust, and therefore use, that system. Explanations and XAI can potentially increase trust,[25] though it is an open question whether they do so in ways that support governance, or whether there are routes other than explainability to build the appropriate trust.[26] Nonetheless, the observations and arguments in this chapter can provide a schema for determining the ways in which explanations of various types do, or do not, support this potential goal of AI governance.

More generally, some high-level observations are in order. First, it is highly unlikely that *any* single AI system could exhibit explainability for all governance goals. Different aspects of AI governance connect with different roles, different goals, and different contexts. The exact same code or algorithm might be appropriate for one role, goal, and context, but not for another. That is, XAI should not really be understood as a type of AI, but rather as a type of AI-individual-society hybrid system, and so efforts at governance via explainability must also have this broader focus. Second, we need to think carefully about whose needs and interests are relevant for a particular governance function, as explanations must be tied to what people actually need and know, rather than an AI researcher's guesses or biases about those. Too frequently, XAI systems are built using the developer's beliefs about what will help deployers, users, or regulators, but without any serious effort to test or confirm those beliefs. One practical response would be to embrace the many calls for increased diversity and participation in the design, development, and deployment of AI systems, as those can help developers better understand the explanation needs of others.

Third, and perhaps most challenging, there is a deep tension between a system being widely interpretable, and it being appropriately governed. The human-interpretable type of XAI—observers can generate a story—is increasingly widespread, particularly through anthropomorphic representations of AI systems. For example, digital assistants (e.g., Siri, Alexa) are designed to help users generate stories about what those assistants "know" or "want," even though those stories are often incorrect. These systems can be interpreted, and "explanations" generated about them, even by people who have no technological training. This kind of XAI is thus particularly appealing for technologies that will be widely deployed, particularly since people arguably need understanding to freely consent to using such systems. However, the previous section showed the ways in which human-interpretable AI is less appropriate than the other two at supporting a wide range of governance functions. Since the human-generated stories need not be grounded in the underlying mechanisms or informational-causal structure of the AI system, they will inevitably fall short for those governance functions that depend on deeper understanding. There is thus an important, unresolved tension that will need to be resolved in coming years: widespread explainability (i.e., most people can generate a story) is insufficient for widespread governance (i.e., systems that are widely deployed and used), but we ultimately require both.

Acknowledgments

Thanks to Cameron Buckner, Jon Herington, Johannes Himmelreich, Ted Lechterman, Juri Viehoff, and Kate Vredenburgh for valuable feedback on earlier versions of this chapter. Significant portions of this chapter were written while the author was on the faculty at Carnegie Mellon University.

[1] Though we should be thoughtful about whether we are, or should be, requiring more from our AI systems than we expect from human decision-makers; see Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: is there a double standard? *Philosophy & Technology, 32*(4), 661-683.
[2] Salmon, W. C. (2006). *Four decades of scientific explanation*. University of Pittsburgh Press.
Van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.
[3] Hempel, C. G. & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science, 15*, 135-175.
[4] Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition, 99*, 167-204.
[5] Some notable early papers include:
Clancey, W. J. (1983). The epistemology of a rule-based expert system: A framework for explanation. *Artificial Intelligence, 20*, 215-251.
Scott, A. C., Clancey, W. J., Davis, R., & Shortliffe, E. H. (1977). Explanation Capabilities of Production-Based Consultation Systems. *American Journal of Computational Linguistics*, 1-50.
Swartout, W. R. (1983). XPLAIN: A system for creating and explaining expert consulting programs. *Artificial intelligence*, *21*(3), 285-325.
[6] There are *many* different introductions and overview of XAI. Some high-level surveys include:
Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE access*, *6*, 52138-52160.
Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, *4*(37).
[7] Biran, O., & Cotton, C. (2017, August). Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)* (Vol. 8, No. 1, pp. 8-13).
Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., & Turini, F. (2019). Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, *34*(6), 14-23.
[8] Hoffman, R., Miller, T., Mueller, S. T., Klein, G., & Clancey, W. J. (2018). Explaining explanation, part 4: A deep dive on deep nets. *IEEE Intelligent Systems*, *33*(3), 87-95.
[9] Examples include:
Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 4768-4777).
Plumb, G., Molitor, D., & Talwalkar, A. (2018, December). Model agnostic supervised local explanations. In *Proceedings of the 32nd international conference on neural information processing systems* (pp. 2520-2529).
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).
[10] Thellman, S., Silvervarg, A., & Ziemke, T. (2017). Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots. *Frontiers in Psychology*, *8*, 1962.
[11] Hempel, C. G. & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science, 15*, 135-175.
[12] Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.
Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.
[13] Friedman, M. (1974). Explanation and scientific understanding. *The Journal of Philosophy*, *71*(1), 5-19.
Kitcher, P. (1981). Explanatory unification. *Philosophy of science*, *48*(4), 507-531.
Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher & W. C. Salmon (Eds.), *Scientific explanation*. Minneapolis, MN: University of Minnesota Press.
[14] Achinstein, P. (1983). *The nature of explanation*. New York: Oxford University Press.
Van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.
[15] Bokulich, A. (2011). How scientific models can explain. *Synthese*, *180*(1), 33-45.
[16] Potochnik, A. (2016). Scientific explanation: Putting communication first. *Philosophy of Science*, *83*(5), 721-732.
[17] Danks, D. (2015). Goal-dependence in (scientific) ontology. *Synthese, 192*, 3601-3616.
[18] Lombrozo, T. (2006). The structure and function of explanations. *Trends in cognitive sciences*, *10*(10), 464-470.
Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*(2), 167-204.
[19] For an overview, see Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, *20*(10), 748-759.

[20] See also Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., ... & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, *296*, 103473.
[21] Lechterman, T. M. (this volume). The concept of accountability in AI ethics and governance. *Oxford Handbook on AI Governance*.
[22] A related question is asked by Zednik, C. (2019). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology*, 1-24.
[23] Though one must be careful to ensure that the proposed actions are actually feasible; see Barocas, S., Selbst, A. D., & Raghavan, M. (2020, January). The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 80-89).
[24] This connection between explainability and ability to determine which failures are "reasonable" has also been discussed by:
Buckner, C. (2020). Understanding adversarial examples requires a theory of artefacts for deep learning. *Nature Machine Intelligence*, *2*(12), 731-736.
Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science, 87*(4), 568-589.
[25] Pu, P., & Chen, L. (2007). Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems*, *20*(6), 542-556.
Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 295-305).
[26] London, A. J. (2019). Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Center Report*, *49*(1), 15-21.