Explaining Explainable AI

Richard Zuroff and Nicolas Chapados

Introduction

An algorithm is a set of steps or instructions to solve a problem. Earlier generations of AI algorithms explicitly codified human expertise to solve problems. For example, "expert systems" developed in the 1970s and 1980s for medical diagnosis captured the knowledge of practicing clinicians in knowledge bases made up of inference rules that predicted the most likely disease diagnosis based on a patient's symptoms. This type of knowledge representation, and associated inference algorithms, allowed any trained physician to fully understand and trace the logic of any diagnosis outputted by the system.[1] In contrast, machine learning approaches extract algorithms directly from data — often by learning useful features[2] and the best ways to combine these features to solve the task. This is a powerful and useful approach, especially when it is difficult or impossible for experts to articulate what the features or rules for combining them should be. However, there is no guarantee that people might, at least in theory, always be able to understand and trace the logic of algorithms developed in this way. This is the challenge of making modern AI systems explainable.

In the last five years, machine-learning-based systems began to be deployed in a wider variety of public and private-sector settings to provide predictions[3] that influence important decisions. As companies' use of AI increased, there was a growing recognition of the importance of transparency due to the risks posed by AI models,[4] and that the so-called "black-box" nature of AI was a critical blocker to broader adoption of the technology.[5] Many researchers responded to these challenges by developing a large number of different techniques to better understand what complex models were actually learning and make their internal processing less opaque. The number of papers with "explainable" or "interpretable" AI in the title or keywords grew more than five-fold from 2014 to 2019,[6] and many review

---

[1] An early classical system was MYCIN, outlined in Buchanan, B.G. & Shortliffe, E.H. (1984). *Rule Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Reading, MA: Addison-Wesley. ISBN 978-0-201-10172-0. An influential follower to MYCIN is CADUCEUS, described in H. Pople. "Evolution of an expert system: from internist to CADUCEUS." Artificial Intelligence in Medicine (1985):179-208.

[2] A feature is an input to the prediction or decision function computed by an AI model and that results from a known (and usually fixed) transformation of one or more raw observed variables.

[3] Prediction is used throughout this chapter as a general term to encompass outputs from AI systems that can include categorization, recommendation, action selection, language generation and more.

[4] For example, as early as 2017, three-quarters of CEO respondents to a PwC survey said the potential for bias and a lack of transparency impede AI adoption in their organizations. See "Explainable AI (XAI)—A Critical Prerequisite For AI Adoption And Success" Forbes. Online: <https://www.forbes.com/sites/intelai/2019/03/27/explainable-aia-critical-prerequisite-for-ai-adoption-and-success/#6a76d4ff3701>.

[5] See for example, McKinsey Quarterly "What AI can and can't do (yet) for your business." https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/what-ai-can-and-cant-do-yet-for-your-business.

[6] Alejandro Barredo Arrieta et al, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI" 2020 Information Fusion, Volume 58 Science Direct online: <https://doi.org/10.1016/j.inffus.2019.12.012>.

articles were published attempting to classify different categories of mathematical techniques that had been developed to probe the inner workings of models.[7] In fact, the explosive growth in technical approaches to explainability outpaced research on the user experience. Many techniques were (and are) published without being tested on human subjects to assess if the explanations that are generated translate to actual changes in variables like increased trust, knowledge, or efficacy.[8]

Perhaps partially as a result of this mismatch, academic interest in explainable AI has only partially translated into more transparent commercial AI systems.[9] Therefore, to ensure the analysis and recommendations in this chapter are as practical as possible for the financial services industry, we adopt a human-centric approach. We do this by framing an explanation as a social interaction between an *explainer* (usually the AI system) and the *explainee* (usually a user or other human stakeholder).[10] We also analyze AI models not as disembodied computer code, but as "sociotechnical systems,"[11] that involve interactions between people, technology, and social structures (such as companies and business processes.) To do that, this chapter incorporates perspectives from other disciplines — such as philosophy, psychology, and law — which have also engaged with fundamental questions of how intelligent agents can and should share information.

Financial institutions have a greater need for explainable AI than many other sectors for several reasons. Partly, this is historical and a result of regulation. For instance, in the U.S., financial institutions must provide clear reason codes when an applicant is refused credit,[12] whether the decision was based on a rules-based system or modern AI. The need for explainable AI also comes from the central role that risk management plays in the financial sector. To develop, use, and supervise AI systems effectively and responsibly, stakeholders need to understand[13] aspects of data collection, data processing, model building, model usage, and more. Third, the use of AI for financial decisions often has important real-world

---

[7] For one literature review see Shane T. Mueller et al, "Explanation in Human-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI" DARPA XAI program, online: <https://arxiv.org/pdf/1902.01876.pdf>.

[8] Erico Tjoa and Cuntai Guan, "A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI" online: <https://arxiv.org/pdf/1907.07374v4.pdf>.

[9] For instance, one study in 2020 found that only one-third of commercial AI products provided explanations for individual decisions. See Vera Liao, Daniel Gruen & Sarah Miller "Questioning the AI: Informing Design Practices for Explainable AI User Experiences" online: <https://arxiv.org/pdf/2001.02478.pdf>.

[10] Tim Miller. "Explanation in Artificial Intelligence: Insights from the Social Sciences" Artif. Intell., 2019

[11] The term is from Miles Brundage et al "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims" https://arxiv.org/pdf/2004.07213.pdf at 5. See also Mark Sendark et al. "The Human Body is a Black Box: Supporting Clinical Decision-Making with Deep Learning" online: <https://arxiv.org/abs/1911.08089> for a discussion of the importance of the sociotechnical framing for real-world applications.

[12] The Equal Credit Opportunity Act (ECOA) is designed to create transparency in the credit underwriting process and protect consumers and business against potential discrimination by requiring creditors to explain the reasons why an adverse action (such as declining a request for credit) was taken. An adverse action must be accompanied by either a statement of the specific reasons for the action taken or a disclosure of the applicant's right to a statement of specific reasons. See Sarah Ammermann, "Adverse Action Notice Requirements Under the ECOA and the FCRA" Consumer Compliance Outlook (2013) online: <https://consumercomplianceoutlook.org/2013/second-quarter/adverse-action-notice-requirements-under-ecoa-fcra/>. Examples of appropriate notifications that show sample reason codes can be found in appendix to Regulation B of ECOA online at: <https://www.consumerfinance.gov/rules-policy/regulations/1002/C/#4>.

[13] Arrieta et al *supra* note 6 use the term "collective sense-making".

economic consequences at the level of consumers, businesses, and the economy as a whole, so it is especially important that these decisions be comprehensible. However, if it is clear that financial institutions need explainable AI, it may be less clear how to operationalize this goal. What does it mean to explain the various elements of an AI system? Who needs to understand which parts? And how can explanations provide that understanding?

The goal of this chapter is to help stakeholders within and outside of financial institutions to answer these questions. The first section unpacks the different types of explanations in terms of the multiple goals that an explainee might have, and highlights how the associated explanation generation techniques may differ. The second section reviews the types of stakeholders who might want explanations, and the legal obligations, risk management concerns, and business imperatives these stakeholders might have that create the need for explanations. The third section identifies the qualities of "good" explanations, and the fourth section identifies why it may be challenging to meet these desiderata. The final section provides recommendations to AI practitioners, risk managers, and senior management on how to develop explainable AI systems that are more likely to be valuable, compliant, and trustworthy.

1. What is an explanation?

There are nearly as many definitions of explanations as articles about the topic.[14] Philosophers have framed explanation as a form of deductive inference, a form of causal reasoning, or a form of statistical inference.[15] To adopt a human-centric approach, this chapter defines an explanation as "a reason for a belief or an action." This definition covers a wide variety of situations and stakeholders for AI systems. For instance, a risk manager might ask why a data scientist *believes* that a model will work well once it is deployed, and a user might ask why they should take the action of *accepting* an algorithm's recommendation.

In fact, there are three different types of questions that the explainee might have, which correspond to three different types of reasons. Internal reasons[16] are why an agent did (or did not do) something based on their beliefs and goals. External reasons[17] ask why an action occurred — independent of the reasons that motivated the agent. Normative reasons ask why something is a good idea or the right thing to do.

1.1. Explanations of internal reasons

Sometimes when an explainee asks for an explanation of an outcome they want to understand what were the *beliefs* and *desires* that lead to it. We refer to this as asking for an explanation about internal reasons. For example, the reason that someone turned their leftovers into lunch for the next day could be that they believed they would not have enough

[14] For a large list of defined terms from the literature see Giulia Vilone & Luca Longo, "Explainable Artificial Intelligence: a Systematic Review" online: <https://arxiv.org/abs/2006.00093> at 9.
[15] Shane T. Mueller et al, *supra* note 10 at 9.
[16] Internal reasons are usually referred to as "motivating reasons" or "agential reasons" in philosophy.
[17] External reasons are referred to as "explanatory reasons" in the philosophical literature. The nomenclature has been changed since this would be confusing in the current context.

time the next morning to prepare something (belief), and they wanted to avoid spending the money to buy lunch (desire). In theory, there are potentially a nearly infinite number of these beliefs and desires: the cook also needed to believe that the leftovers would still be edible the next day, and that no one would steal them from the fridge overnight, and that they would still want to eat food at lunchtime.[18] In practice, people select a small number of reasons to be the explanation,[19] a point we return to in section 3 on what makes for good explanations.

Even today's most advanced AI systems do not have the consciousness required to have desires, but it is fairly uncontroversial to claim they have close analogues in the form of goals. So in the realm of AI, one category of explanation is reasons why the software or artificial agent took some action (such as making a particular prediction) based on its information, beliefs,[20] and goals. Explanations that provide the relative importance of variables — such as using payment and credit usage patterns and ratios like age-to-income for predicting loan defaults — are one example of internal explanations. They explain an outcome (such as a recommendation that a person's application for credit be declined) based on the AI system's information (such as their income and debt ratios) and goals (such as meeting business goals and respecting the financial institution's risk appetite).

Three important categories of techniques to generate internal explanations are feature contributions, examples, and counterfactuals. Shapley values,[21] among the oldest and most used algorithmic techniques in the field, provide insight into a system's reasoning by evaluating the relative contribution of each feature to a model's decision. This is done by drawing from game theory and testing various groups (or "coalitions") of features and evaluating the importance of individual features for these groups. LIME,[22] arguably the second best-known technique of the field, also provides evaluations of the importance of features but by building and then perturbing an interpretable — locally linear — surrogate to the original model. Also in the class of feature contribution methods are activation atlases, which, rather than probing a surrogate, show how different parts of the original model respond more or less strongly to different inputs.[23] This is somewhat analogous to how fMRI studies examine how different parts of the brain light up under stimuli.

Another internally focused category of explanation-generation techniques is to provide explainees with quintessential examples of each category. This allows people to apply case-based reasoning to understand why similar examples had similar outcomes.[24] Showing criticisms — examples that break with prototypical expectations — also helps

[18] Maria Alvarez, "Reasons for Action: Justification, Motivation, Explanation", The Stanford Encyclopedia of Philosophy (Winter 2017 Edition), Edward N. Zalta (ed.), online: <https://plato.stanford.edu/archives/win2017/entries/reasons-just-vs-expl/>.
[19] Miller, *supra* note 10 at 6.
[20] In Bayesian reasoning methods a probability distribution is generally interpreted as a "degree of belief" (prior or posterior to an observation).
[21] Lundberg, Scott M., and Su-In Lee. 2017. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems, *30*, 4765-4774
[22] Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin "'Why Should I Trust You?' Explaining the Predictions of Any Classifier" online: <https://arxiv.org/abs/1602.04938>.
[23] Shan Carter et al. "Exploring Neural Networks with Activation Atlases" online: <https://distill.pub/2019/activation-atlas/>.
[24] See Been Kim, Cynthia Rudin and Julie Shah, "The Bayesian Case Model: A Generative Approach for Case-Based Reasoning and Prototype Classification" online: <https://arxiv.org/abs/1503.01161>.

understanding.[25] A further extension of this idea is to explain a decision or prediction based on comparison with prototypical parts. This is easiest to imagine in the visual realm: for instance an algorithm could explain its classification of a bird into a specific species by showing various training images showing a prototypical beak, belly, and feathers.[26] An equivalent in the realm of finance might be an algorithm that explains its forecasts of cash flows by showing prototypical time series segments for different seasonal effects, such as strong holiday sales, post-holiday returns, and quieter times of year. Counterfactual techniques help end users understand more about the internal reasons for a specific prediction by explaining how the answer would have changed if some aspect of the input had been different. For instance, a counterfactual explanation of a loan decline decision might be that the applicant's debt ratio was too high and they would have been approved if the ratio was lower. This explanation is focused on an aspect of the input to a model rather than its processing logic, but is still focused on the internals of a system.

1.2. Explanations of external reasons

People sometimes have mistaken beliefs, which could make internal explanations unsatisfying. For example, imagine that person A mistakenly believes that person B betrayed them. If person A gets angry at person B and someone asks "why?", it seems wrong to say that it is *because of* an event (B's betrayal) that did not actually happen. In this case the mistake — the mismatch between person A's beliefs and the truth — is a better explanation. We refer to cases like this, where an explainee is looking for causes that are outside of internal states to understand an error, as external reasons. It is an interesting question whether people can be mistaken about their own goals,[27] but there are many examples of algorithms mistaking their designers' goals.[28] Put differently, it is an ongoing challenge for designers to fully and correctly align AI's goals with their own intentions,[29] which may be informal or at least less than perfectly mathematically precise. Consequently, there are many examples of algorithms surprising their creators, such as a game-playing AI that learned to repeatedly crash a boat rather than completing a race.[30] A physical (but hypothetical) example would be a warehouse robot that learns to more quickly finish its task by throwing

---

[25] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. "Examples are not enough, learn to criticize! criticism for interpretability." In Advances in Neural Information Processing Systems*, 29*, 2280–2288.
[26] Chaofan Chen, "This Looks Like That: Deep Learning for Interpretable Image Recognition" online: <https://arxiv.org/abs/1806.10574>.
[27] Ian Deweese-Boyd, "Self-Deception", The Stanford Encyclopedia of Philosophy (Fall 2017 Edition), Edward N. Zalta (ed.), online: <https://plato.stanford.edu/archives/fall2017/entries/self-deception/>.
[28] See Victoria Krakovna, "A master list of examples of AI systems gaming their objective specification" online:
<https://docs.google.com/spreadsheets/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bjOxCG84dAg/pubhtml>
[29] In particular, see D'Amour, Alexander, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen et al. (2020) "Underspecification presents challenges for credibility in modern machine learning." arXiv preprint arXiv:2011.03395.
[30] Open AI, "Faulty Reward Functions in the Wild" online: <https://openai.com/blog/faulty-reward-functions/>.

packages into the garbage rather than a truck.[31] Asking why it failed to properly load the truck is a request for an external explanation that goes beyond the model's internal states.

A real-world example of this type of external explanation can be found in the analysis of an early medical AI system for evaluating the risk of patients with pneumonia. The developers found that the system often recommended that asthmatics be given lower priority for treatment based on their higher-than-average survival rates.[32] Any human doctor would see this as an obvious mistake. Asthmatics with pneumonia are at high risk, and only had high survival rates because standard healthcare procedure was to immediately provide them with intensive care.[33] An external explanation of why an AI model recommended an asthmatic with pneumonia be deprioritized for treatment would need to include information outside of the algorithm, such as what data was (or was not) collected and what assumptions were in place. Similarly, when systems show surprising behavior, an external explanation could be that the design failed to capture something important to the AI system designer. For instance, a trading algorithm might learn that placing and cancelling orders was an effective way to learn about demand for different securities to find the best venue for execution, or with the goal of actively changing demand and thus price. Under current U.S. law, the latter might qualify as illegal spoofing while the former is acceptable market-making.[34] When a regulatory body asks why the AI system was placing and cancelling orders, the explanation might have to come from an analysis that steps outside of the algorithm to consider potential issues with the data, design, or implementation.

An important category of external explanation-generation techniques is traceability tools that describe and track how a system was trained, how it was evaluated and tested, and how it is monitored. A basic version of this known as model cards. A model card captures information such as the origins of the training data and quality checks that were applied; how the model was evaluated and why evaluation datasets were representative; limitations of the testing and uses of the model that should be avoided due to those limitations.[35] External errors can often be due to interactions between the different portions of an AI system. By chaining or reusing models' outputs, systems may develop hidden dependencies and build-in hidden assumptions or other forms of hidden technical debt[36] that make AI systems less reliable or

---

[31] For example, in one study of Lego stacking, the desired outcome was for a red block to end up on top of a blue block. Instead of picking up the red block and placing it on top of the blue one, the agent simply flipped over the red block because the reward function was based on the height of the bottom face of the red block. Deep Mind. "Specification gaming: the flip side of AI ingenuity" online: <https://deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity>.

[32] Cliff Kuang, "Can A.I. Be Taught to Explain Itself?," The New York Times, November 21, 2017, sec. Magazine, https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html. Original paper: <http://people.dbmi.columbia.edu/noemie/papers/15kdd.pdf>.

[33] In essence the model understood the causal influence backwards because it was built to extract correlations from historical data. How to build models that make correct causal inferences from purely observational data is an active topic of research. See for example: Stephen L. Morgan & Christopher Winship. 2014. *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* Cambridge: Cambridge University Press.

[34] United States v. Coscia, No. 16-3017, 2017 WL 3381433 (7th Cir. Aug. 7, 2017). online: <http://media.ca7.uscourts.gov/cgi-bin/rssExec.pl?Submit=Display&Path=Y2017/D08-07/C:16-3017:J:Ripple:aut:T:fnOp:N:2006533:S:0>.

[35] See Margaret Mitchell "Model Cards for Model Reporting" online: <https://arxiv.org/pdf/1810.03993.pdf>.

[36] D. Sculley et al, "Hidden Technical Debt in Machine Learning Systems" Advances in Neural Information Processing Systems 28 (NIPS 2015) online: <https://papers.nips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf.

reproducible. As a result, the idea of model cards has been generalized to cover the whole system of machine learning systems in the framework of Technological Readiness Levels.[37] Technical Readiness Levels are meant to be a principled process to ensure robust systems, inspired by practices in safety-critical contexts such as spacecraft systems, which could support the creation of external explanations.

Explanations that surface relevant examples can also offer valuable external explanations. Highlighting examples that were the most influential to a model's incorrect prediction can suggest external reasons, such as poorly chosen or mislabeled samples in the training dataset.[38] Counterfactuals can be useful for finding potential external reasons. However, unlike counterfactuals used for internal reasons, these should be applied in a directed sense to allow investigative explanations. For instance, a counterfactual explanation around loan approval could be generated to evaluate a model's reaction to specific low and high values of certain features and these explanations would hint at both internal and external reasons that are pushing the model to this decision.

1.3. Explanations of normative reasons

The term "normative reason" comes from the notion that there are norms, principles, or codes that make it right or wrong to do certain things. These could be ethics, or binding rules such as laws, or even just specific customs of a group of people. Normative reasons are often referred to as justifications**,** and are different from internal (motivating) reasons. For example, a city's mayor might ban the sale of sugary drinks because it is the right thing to do in order to protect citizens from diabetes and other diseases (the normative reason). She also might institute a ban because she believed it would be popular among voters and she wanted to be re-elected (the internal reason). Both explanations can be true, but the normative and internal reasons are different. (In practice, part of the role of administrative law in many jurisdictions is to ensure that the normative reasons given by public or semi-public institutions for their decisions are aligned with their internal reasons.)
Currently, algorithmic decisions are generally held to the same standard as humans.[39] Early misconceptions that decision-making by algorithms was inherently objective and therefore fair has given way to growing recognition from policymakers, regulators, and advocates that AI raises novel challenges for ensuring non-discrimination, due process, and that other

[37] A. Lavin, et al. (2021). "Technology readiness levels for machine learning systems" arXiv preprint arXiv:2101.03989.
[38] Elnaz Barshan, Marc-Etienne Brunet, Gintare Karolina Dziugaite, "RelatIF: Identifying Explanatory Training Examples via Relative Influence" online: <https://arxiv.org/abs/2003.11630>.
[39] For instance, there is strong policy support, at least in Europe, for the "principle of functional equivalence"which holds that for the purposes of assessing harm to third parties, "using the assistance of a self-learning and autonomous machine should not be treated differently from employing a human auxiliary." See Expert Group On Liability And New Technologies – New Technologies Formation, "Liability for Artificial Intelligence and other Emerging Digital Technologies" (2019) online: <https://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupMeetingDoc&docid=36608> at at 25. However it has been proposed that in some situations AI systems can and should be held to a higher standard than human decision-makers. See for example, Finale Doshi-Velez et al, "Accountability of AI Under the Law: The Role of Explanation" online: <https://arxiv.org/abs/1711.01134> at 17.

norms such as transparency and accountability are upheld.[40] For example: one survey identified 23 different potential sources for bias in AI systems ranging from relatively obvious technical issues (such as non-random sampling across demographic groups) to subtler issues (such historical real-world bias that can seep into the representations used by intermediate processing steps of the algorithm even with perfect sampling).[41] When an explainee asks why a decision is justifiable or fair for her, she is looking for normative reasons, which requires looking both within and outside of the decision-making algorithm itself. For example, when an algorithm recommends than an applicant's loan be denied, statistics that show the likelihood of an error across race, gender, or other protected attributes could be a type of normative explanation.

However, in this example, the decision-subject or other explainee (such as a regulator) is unlikely to find the statistics to be a compelling normative explanation unless they are convinced that the measure properly reifies the underlying norm of "fairness". There are many different potential measures of fairness[42] — and every other important norm. As a result, it is difficult to create automated, mathematical techniques for generating normative explanations as the very nature of normative reasons resist simple encodings. For example, it might be tempting to consider a technique that identifies whether a variable like race or gender or close proxies to these types of sensitive attributes is present in a model to be a good normative explanation. However, the presence or absence of variables in a model does not indicate the presence or absence of a causal connection,[43] so assessing whether there was any discriminatory impact requires a more sensitive analysis. In fact, sensitive variables can be included in a model precisely to impose a measure of fairness into a model's logic. Showing that fairness constraints[44] (or other normative constraints) have been imposed can be an effective type of normative explanation. Similarly, counterfactual techniques can also be used to assess whether a decision was truly "because of" a protected or sensitive attribute.[45]

The examples above focused on decision-subjects and regulators. Business users of AI systems may also be interested in normative explanations. For instance, many financial professionals are CFA charterholders, who have ethical obligations such as ensuring their advice is suitable for clients.[46] They may want to ensure they continue to uphold these norms

[40] See for example, Frederik Zuiderveen Borgesius "Discrimination, artificial intelligence, and algorithmic decision-making" Council of Europe online: <https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73> and Osonde A. Osoba et. al, "Algorithmic Equity A Framework for Social Applications" RAND online: <https://www.rand.org/content/dam/rand/pubs/research_reports/RR2700/RR2708/RAND_RR2708.pdf> at 18.
[41] Ninareh Mehrabi et al "A Survey on Bias and Fairness in Machine Learning" online: <https://arxiv.org/pdf/1908.09635.pdf>.
[42] See Ninareh Mehrabi et al, "A Survey on Bias and Fairness in Machine Learning" online: <https://arxiv.org/abs/1908.09635> at 11 for one list.
[43] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. "Counterfactual fairness". In Advances in Neural Information Processing Systems, pages 4066–4076, 2017.
[44] See: Muhammad Bilal Zafar et al "Fairness Constraints: A Flexible Approach for Fair Classification" as well as Louizos, C., Swersky, K., Li, Y., Welling, M., & Zemel, R. S. (2016). The Variational Fair Autoencoder. In Proceedings of the International Conference on Learning Representations (ICLR).
[45] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva *supra* note 43.
[46] CFA Institute, "Code Of Ethics And Standards Of Professional Conduct" online: <https://www.cfainstitute.org/-/media/documents/code/code-ethics-standards/code-of-ethics-standards-professional-conduct.ashx>.

even if they are using algorithms to help make recommendations, such as by asking for explanations of how the algorithm is taking into account different clients' risk tolerances. Managers who fail to ask about whether AI systems have been properly trained to achieve norms such as suitability might be creating legal risks related to negligence (as discussed in section 2.4), and thus also be interested in normative explanations.

1.4. Types of explainability technique

There are many specific techniques that have been developed to answer these different desires for reasons, and any attempt to comprehensively list explainability techniques will likely be outdated soon after it is published. Researchers have developed explanation-generation techniques that vary depending on the model being explained (such as neural networks, Bayesian networks, expert systems, or model-agnostic approaches), the type of input data and explanation output format (numerical, textual, visual, rule-based, or mixed format), what type of problem the model is being used to solve (such as classification or regression), and whether the technique is applied during or after training. A partial list is in the table below.[47]

Output type *Input type*	**Numerical explanations**	**Rules-based explanations**	**Visual explanations**	**Textual explanations**	**Mixed explanations**	**More interpretable surrogate model**
Numerical/ categorical	Causal Importance, Influence Functions	Neural Network Knowledge eXtraction, Tree Regularization	t-SNE maps, ActiVis, Deep Learning Important FeaTures	Contextual Importance, Utility, Causal Importance	Deterministic Finite Automaton	General Additive Models, Interpretable Classification Rule Mining
Visual	Maximum Frequency Difference Probes, Concept Activation Vectors	DT Extraction, Discretized Interpretable Multi Layer Perceptrons	Saliency maps, Activation maps, Receptive Fields,	Relevance, Discriminative Loss Seq2seq-Vis	Attention Alignment, Maximum Mean Discrepancy -critic, Image Caption Generation with Attention Mechanism	SHapley Additive exPlanation, Argumentation
Textual	Influence Functions, REcurrent LEXicon NETwork	Word Importance Score, Iterative Rule Knowledge Distillation	Seq2seq-Vis, LSTMVis, SHAP	REcurrent LEXicon NETwork, Rationales	Evasion-Prone Samples Selection, Local Interpretable Model-Agnostic Explanations	Unsupervised Interpretable Word Sense Disambiguation
Time series		Validity Interval Analysis, Anchors		Most-Weighted-Path, Most Weighted-Combination		Fuzzy Inference Systems, Gaussian Process Regression

Note that, for simplicity, this table does not distinguish between different strategies employed to generate explanations. A technique can try to separate and quantify the influence of its features (separately or in combinations), or construct a simpler model that

[47] The framework nomenclature and selected examples are from a larger list in Giulia Vilone & Luca Longo, *supra* note 14.

approximates the workings of the complex model overall (i.e., globally, for all the input data). Some techniques enable developers to construct several simpler models that each approximate parts of the more complex model (such as subsections of the data or input space) while others aim to explain individual predictions rather than attempting to create any kind of a surrogate for the whole or even parts of the complex model.

Another critical distinction in considering different explanation generation techniques is whether the result will be an AI system that is *transparent*, *interpretable*, or *comprehensible*. In transparent systems, the mappings from inputs to outputs are visible to an investigator even if they cannot understand anything about the mapping. For example, alongside one of its predictions, a deep neural network might also provide a file that lists every one of the millions of connection weights between its nodes, as well as the activation values of the nodes when the prediction was made. This is fully transparent, but not something any human could interpret in its raw form. Nonetheless, this type of file might be useful to a regulatory agency with some specifically designed software to process the information.

In interpretable systems, investigators can not only see, but also study and understand how inputs are mathematically mapped to outputs. These mappings may not correspond to any natural human concepts, but they still allow the algorithm's logic to be probed and explored. For example, a machine vision model might highlight the pixels and portions of an image that were most influential in its decision. Some of these saliency maps might correspond to recognizable things or scene features, but many others may be incomprehensible to a human observer.[48] That is, there may be untranslatable 'words' in the machine's 'language' in the same way that some cultural-specific words may be difficult to translate into other human languages.[49] For instance, a trading algorithm might use some recognizable features (such as "liquidity" or "ratio of short- to long-term traders") to plan an execution strategy, but other features may be mathematical abstractions with no clear human analogue.[50]

To be a comprehensible system, investigators must be able to map the relationships between inputs and outputs using concepts (perhaps such as "credit risk", "liquidity risk", "duration risk" or "economic cycles") that correspond to the investigator's existing ways of structuring knowledge about the world. Ordinary users such as consumers would have the most obvious

---

[48] For example one study compared where humans and deep networks looked in scenes to answer questions. In some cases attention was similar (such as looking at the region around a cat's eye to answer "what colour is the cat's eye") in other cases the network's attention map diverged significantly from humans (such as looking at a person's shoulder, foot, and background rather than a a frisbee in flight to answer "what is the man doing"). Abhishek Das et al. "Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions?" https://arxiv.org/pdf/1606.03556.pdf

[49] The analogy comes from Been Kim https://www.youtube.com/watch?v=jXkhcoZ9W24. Sometimes this is literally true as the intermediate representations generated by machine translation algorithms do not correspond to any human language. See Melvin Johnson et al. "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation." Transactions of the Association for Computational Linguistics" , [S.l.], v. 5, p. 339-351, oct. 2017.

[50] This is not solely an issue with more complex model architecture such as neural networks, linear systems are not necessarily more interpretable. In many domains, in order to get comparable performance to neural networks, linear models must operate on heavily hand-engineered features that are themselves fairly inscrutable. See Zachary C. Lipton, David C. Kale & Randall Wetzel, "Modeling missing data in clinical time series with rnns." Machine Learning for Healthcare, 2016, cited in Zachary C. Lipton, "The Mythos of Model Interpretability" online: <https://arxiv.org/abs/1606.03490>.

need for comprehensible (not just interpretable) systems, but what counts as recognizable features may be different for groups with expert knowledge. For instance, one hedge fund built and tested models based on data from its portfolio managers to try and mimic their trades.[51] Because the model was optimized to replicate traders' patterns (rather than to predict stocks that would outperform), there is a higher likelihood that the underlying factors (such as trade timing, market liquidity, and the duration over which managers built positions) will be comprehensible to an investing professional. For users with less specialized expertise it may be useful to probe whether the model is using an ordinary mental concept (such as "stripes") in its internal logic to assess its comprehensibility.[52]

This is one example of the larger point that, because the benefits of explainability are mediated by the reactions of the people who receive them, it is important for financial institutions to clearly understand the different personas who might be requesting and receiving explanations. The needs of different stakeholders are explored in more detail in the next section.

2. Why are explanations needed?

There are at least nine different types of stakeholders within and outside of a financial institution who might want explanations. These are:

1. the subjects of algorithmic decisions (such as ordinary consumers);
2. the data scientists and designers who develop the model, system, and interfaces;
3. business users of the outputs of an algorithm (such as front-office employees);
4. managers and executives who are responsible for the business outcomes and risks tied to use of the model;
5. model checkers within the business (such as other data scientists or an analytics center of excellence) who act as the first line of defense[53] in risk management;
6. independent model checkers in risk and compliance functions who act as the second line of defense for risk management;
7. external auditors who act the third line of risk management defense;
8. conduct regulators focused on enforcing specific rules (such as fair lending or securities laws);
9. prudential regulators focused on the safety and soundness of the overall financial sector.[54]

---

[51] Saijel Kishan, "Steve Cohen Is Trying to Teach Computers to Think Like Top Traders" March 15, 2017 Bloomberg online: <https://www.bloomberg.com/news/articles/2017-03-15/steve-cohen-said-to-eye-computers-to-model-top-traders-thinking>.

[52] Been Kim et al, " Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)" online: <https://arxiv.org/abs/1711.11279>.

[53] Institute of Internal Auditors, "The Three Lines Of Defense In Effective Risk Management And Control" 2013 IIA Position Paper online: <https://na.theiia.org/standards-guidance/Public%20Documents/PP%20The%20Three%20Lines%20of%20Defense%20in%20Effective%20Risk%20Management%20and%20Control.pdf>.

[54] See Henrike Mueller, Florian Ostmann, "AI transparency in financial services – why, what, who and when?" FCA online: <https://www.fca.org.uk/insight/ai-transparency-financial-services-why-what-who-and-when>.

These different personas will have different motivations for requesting explanations, and different stakeholders may be looking for different levels or types of insight. For instance, a front office employee who is incentivized to drive revenue might be more likely to want an explanation when an algorithm recommends declining or not proceeding with a sale, and a business manager might be interested in the mean and variance of outcomes in order to develop financial and operational plans. In contrast, a risk professional might be most interested in explanations of "why" a transaction should be allowed to proceed rather than "why not". For them, explanations are needed to help calibrate trust and manage risk.

Commentaries on explainability often focus on one or a few specific motivations for providing explanations, often the direct obligation to provide explanations to consumers created by some legal regimes. In fact, for financial institutions, these direct obligations are likely to be only a small part of the legal, operational, and reputational motivations for generating explanations. Financial institutions deploying AI should take a holistic view of the need for explanations, both to ensure these varied risks are well-managed, but also to maximize the value that can be created with AI systems. In the subsections below, we identify six different motivations or obligations to provide explanations.

2.1. Direct obligations to consumers

One of the earliest sources of rights and obligations related to explanations of algorithmic decision-making is the European Union's General Data Protection Regulation (GDPR), which was adopted in 2016 and became enforceable in 2018.[55] GDPR applies if the data, data subject, data collector, or data processor are based in Europe, and so is relevant to any financial institution with European customers or operations. However GDPR has also become a model for national data protection laws and policy, including with regard to explanations for algorithmic decision-making, outside of the EU, including Singapore,[56] Canada,[57] and Brazil.[58] Other countries, from Colombia to Bermuda, are "adopting European

[55] General Data Protection Regulation, Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, [2016] OJ L 119 at art. 3 [GDPR].

[56] The Personal Data Protection Act 2012 (No. 26 of 2012) Singapore. For a comparison of this act with GDPR see OneTrust DataGuidance "Comparing privacy laws: GDPR v. Singapore's PDPA" online: <https://www.dataguidance.com/sites/default/files/gdpr_v_singapore_final.pdf>. The Monetary Authority of Singapore has also issued non-binding principles for the domestic financial sector. The principles recommend that "Data subjects are provided, upon request, clear explanations on what data is used to make AIDA-driven decisions about the data subject and how the data affects the decision" and that "Data subjects are provided, upon request, clear explanations on the consequences that AIDA-driven decisions may have on them." Monetary AUthority of Singapore, "Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector" online: <https://www.mas.gov.sg/~/media/MAS/News%20and%20Publications/Monographs%20and%20Information%20Papers/FEAT%20Principles%20Final.pdf>.

[57] The government of Canada has proposed amendmending its main privacy legislation to include the right to an explanation explicitly modelled on GDPR. At the time of writing (early 2021), this right had been proposed in Bill C-11 but was not yet effective law. Bill C-11, Digital Charter Implementation Act, 2020, 2nd Session, 43rd Parliament, 2020.

[58] Brazil's Lei Geral de Proteçao de Dados (LGPD) was modeled directly after GDPR and includes a similar obligation to that in the Guidelines: "whenever requested to do so, the controller shall provide clear and adequate information regarding the criteria and procedures used for an automated decision, subject to commercial and industrial secrecy." article 20. Translated text from: <https://iapp.org/media/pdf/resource_center/Brazilian_General_Data_Protection_Law.pdf>.

rules almost word for word,"[59] and several other countries, including Japan and South Korea, either have received or are in the process of securing adequacy decisions that certify their domestic data protection regimes provide the same safeguards as Europe even if they are drafted differently.[60] As a result of GDPR's extraterritorial reach and influence, the large numbers of expatriate citizens around the world, and the difficulty of fragmenting internal processing pipelines to treat data originating in Europe differently, GDPR compliance is likely to be treated as the "gold standard" for data protection for most companies.[61] This implies that all multinational financial institutions, as well as domestically-focused financial institutions in the countries cited above, may have explicit obligations to provide explanations based on GDPR.

At a high-level, the logic of explainability in GDPR is that a) the regulation prohibits automated processing for significant decisions unless there are safeguards in place; b) one such safeguard is the right of data subjects to receive explanations of how and on what basis the decision was made; c) these explanations are prerequisites for subjects to effectively exercise their other rights (such as contesting a decision); and d) explanations are also prerequisites for systemic (or institutional) controls on automated decision-making such as algorithmic auditing.[62] Specifically, GDPR Article 22 states that individuals "shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her" unless there is a suitable basis for processing and there are "suitable measures to safeguard the data subject's rights."[63] The Guidelines accompanying GDPR specify that decisions which "affect someone's financial circumstances" — such as approving or denying credit — would likely qualify as significantly affecting someone,[64] and many other financial products that are not explicitly named in the Guidelines could qualify. GDPR Article 15 also states that if there is solely automated-decision making that falls within Article 22, then the data subjects have the right to access information about "the logic involved, as well as the significance and the envisaged consequences of such processing."[65] Scholars have also argued that a coherent reading of the two articles implies that if "the

[59] See Mark Scott & Laurens Cerulus, "Europe's New Data Protection Rules Export Privacy Standards Worldwide" Politico (Jan. 31, 2018), https://www.politico.eu/article/europe-dataprotection-privacy-standards-gdpr-general-protection-data-regulation/

[60] European Commission. "Adequacy Decisions" Accessed at: https://ec.europa.eu/info/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en

[61] Margot E. Kaminski. "The Right To Explanation, Explained" Berkeley Technology Law Journal, Vol. 34, No. 1, 2019 at 187.

[62] This interpretation relies on both the actual text (Articles) of GDPR, and its preamble (Recitals) and enforcement Guidelines. While only the Articles have direct force of law, the Recitals are cited by courts as authoritative interpretations, and the Guidelines have been promulgated by a group representing the Data Protection Authorities who will actually enforce the law. See *Ibid.* For an earlier contrary view, see Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. "Why a right to explanation of automated decision making does not exist in the general data protection regulation." International Data Privacy Law, 7(2): 76–99, 2017

[63] Regulation (EU) 2016/679, of the European Parliament and the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) 1 at arts. 22,15 [hereinafter *GDPR*].

[64] Data Protection Working Party, "Guidelines on Automated Individual Decision-Making and Profiling for the Purpose of Regulation 2016/679" (2017) WP251rev.01, [*GDPR Guidelines]* at 21.

[65] *GDPR* 15g

weights and factors" used in the decision are necessary to check its accuracy and to potentially challenge its correctness, then such information needs to be part of a valid explanation, not just the general structure of the algorithm.[66] Thus, GDPR (and related laws) create direct obligations to explain the logic, features, and consequences of AI system's decisions.

2.2. Justifying the basis for conduct

While GDPR and related laws apply across industries, legislation specific to the financial sector can also create demand for explanation in order to show that a financial institution's conduct that was based on an AI system had an appropriate basis. For example, in some jurisdictions any investment recommendations and ongoing investment advice must be suitable and appropriate for each client.[67] Robo-advisors provide automated, personalized investment advice and monitoring, and portfolio management activities (such as tax-loss harvesting) to satisfy customers' predefined investment strategies. In the U.S., robo-advisers are subject to the same requirements that their advice must be suitable and appropriate for each client,[68] so explanations might be required to demonstrate suitability. For instance, a robo-advisor might demonstrate that if a customer's risk tolerance or investment horizon had been lower, then the recommendations would have changed.[69]

Another example comes from fair-lending laws. In the U.S., financial institutions must provide clear reason codes when an applicant is refused credit, which are meant to ensure lenders' business practices are not based on protected categories such as race or religion.[70] Therefore, if an AI system is used to evaluate creditworthiness, it must also be able to provide explanations of the basis for adverse decisions.[71] The potential need to explain the basis of AI decision-making can also arise in antitrust and securities law contexts.[72] For example, if a financial institution uses AI to appraise homes that will serve as collateral for mortgage-backed loans which are packaged into mortgage-backed securities, an investor in these securities who loses money might complain or sue under the theory that the valuations were incorrect.[73] The potential plaintiff might argue that the AI ignored important factors,

---

[66] Prior to the Guidance some scholars argued that "the logic involved" only refers to the general structure and architecture of the processing model. See Andrew D Selbst, Julia Powles. "Meaningful information and the right to explanation" International Data Privacy Law, Volume 7, Issue 4, November 2017) referenced in Kaminski *supra* note 61.

[67] See SEC. "Guidance Update February 2017 | No. 2017-02" Accessed at: https://www.sec.gov/investment/im-guidance-2017-02.pdf

[68] *Ibid*.

[69] In re Westmark Financial Services, Corp., Advisers Act Release No. 1117 a financial planner recommended speculative equipment leasing partnerships to unsophisticated investors with modest incomes, referenced in *Ibid.*

[70] US Congress. 1970-10. 15 U.S.C. 1681 Fair Credit Reporting Act. online: <https://www.gpo.gov/fdsys/pkg/STATUTE-84/pdf/STATUTE-84-Pg1114-2.pdf>

[71] See Jiahao Chen. "Fair lending needs explainable models for responsible recommendation." In Proceedings of Workshop on Responsible Recommendation, Vancouver, Canada, October 6, 2018 (FATREC'18), 4

[72] For example, "some courts have held that where the conduct underlying a market manipulation is an open-market transaction, such as a purchase or short sale, there must be a showing that the conduct lacked any legitimate economic reason before there can be liability under Section 10(b)." Yavar Bathaee. "The Artificial Intelligence Black Box And The Failure Of Intent And Causation." Harvard Journal of Law & Technology, Volume 31, Number 2 Spring 2018 at 910.

[73] The example is from *Ibid* at 914.

such as homes' age or square footage. In this and similar situations the issuer would benefit from being able to provide explanations of the system either to deter litigation or defend its conduct. These might be internal explanations, such as a list that ranks the relative importance of features, or more external explanations, such as the procedures that were used to test and validate the models' assumptions and outputs.

2.3. Rebutting negligence claims

The need for explanations to satisfy conduct basis tests will emerge from industry-specific regulation, such as securities law, but the broader legal category of negligence can also create a need for explanations of AI systems. For instance, in many jurisdictions, the law provides causes of action against employers for the negligent hiring, supervision, or retention of human employees as their agents.[74] Scholars have argued that artificial agents (AI systems) should be treated as legal agents for the purposes of negligent supervision claims[75] and the general regulatory principle of functional equivalence suggests that "a person using a technology which has a certain degree of autonomy should not be less accountable for ensuing harm than if said harm had been caused by a human auxiliary."[76]

We believe that courts (at least in the U.S. and Europe) may be open to allowing liability for negligent training of an artificial agent because this incentivizes better care in design and programming for the system developer. While this is speculative today, known failure modes might help to define the standard for negligent training and supervision. AI practitioners have begun to collect cases of undesired or unexpected behavior by AI systems.[77] As these become codified into best practices, financial institutions will likely want to be able to explain how their systems were properly trained in reference to common failure modes in order to rebut claims that any resulting harms were foreseeable. European experts have also recommended imposing a duty to "equip technology with means of recording information about the operation of the technology (logging by design)" and that the absence of logged information should trigger a presumption that the condition of liability has been proven.[78] If this type of recommendation is widely adopted, then financial institutions will want to be able to rebut this presumption by having logging and other system-level explanations of behavior.

It should also be noted that when accidents happen that involve semi-automated systems, there is a tendency to shift the focus from systems to people ("operator error") even when the system is advertised as being highly resilient.[79] Most real-world failures are not due to

---

[74] See for example, Seth E. Lipner & Lisa A. Catalano, The Tort of Giving Negligent Investment Advice, 39 U. MEM. L. REV. 663, 668 (2009).
[75] See Samir Chopra, Laurence F. White. A Legal Theory for Autonomous Artificial Agents at 133 cited in Ignacio N. Cofone, Servers and Waiters: What Matters in the Law of A.I., 21 STAN. TECH. L. REV. 167, 191 (2018)
[76] See Expert Group On Liability And New Technologies *supra* 39 at 3.
[77] Partnership on AI, "AI Incident Registry" PAI online: <http://aiid.partnershiponai.org/>
[78] Expert Group On Liability And New Technologies *supra* note 39 at 7.
[79] Human actors that bear the brunt of moral and legal responsibilities when a semi-autonomous system (like an airplane autopilot) fails have been referred to as "moral crumple zones". Madeleine Clare Elish, "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction" (pre-print) (March 1, 2019). Engaging Science, Technology, and Society (pre-print). Available at SSRN: <https://ssrn.com/abstract=2757236>.

errors purely within either human or machine processing, but result from interactions between the two. For example, the highest risk periods for semi-autonomous vehicles occur when control is transferred from assisted driving technology back to the human,[80] and research has shown the people are often unsure or misunderstand when they can take back control from an automated system.[81] In using explanations to protect against or assess negligence claims, it is important that financial institutions, regulators, and courts consider the true causes of harm, which will often involve complex interactions between users, models, interaction-design, and other factors, rather than focusing explanations solely on the mistakes made by people.

2.4. Algorithmic auditing

Claims based on GDPR-style rights, conduct requirements, and negligence are all forms of individual rights. Some commentators have suggested that individual rights are insufficient on their own to ensure the safe and responsible use of AI.[82] Algorithmic auditing and certification are two proposed protections from automated decision-making that are more institutional (although they may be in service of the individual rights discussed in section 2.1). An algorithmic audit is a "mechanism to check that the engineering processes involved in AI system creation and deployment meet declared ethical expectations and standards, such as organizational AI principles."[83] The goal of an algorithmic audit is to "prove that [machine learning algorithms] are actually performing as intended" and not producing discriminatory, erroneous or unjustified results.[84]

Some degree of transparency, such as an internal explanation, is likely required to assess how an AI algorithm is "actually" performing, and can help uncover accidental discimination or other unjustified results. For instance, in being trained models may learn to use proxies for protected variables such as zip codes proxying for race (or other protected

---

[80] In the case of the first fatal crash involving an autonomous vehicle, while the AI system detected the pedestrian 5.6 seconds before impact, it never accurately classified her as a pedestrian. By the time the system determined that a collision was imminent, the required response time was too rapid for the automated braking system so it relied on the operator. However the operator was looking at her phone and only redirected her gaze to the road 1 second before impact, and she did not have enough time to finish steering away from the pedestrian. The National Transport Safety Board investigation concluded that the probable cause of the crash was the failure of the vehicle operator to monitor the driving environment and the operation of the automated driving system because. See National Transport Safety Board "Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian Tempe, Arizona March 18, 2018" Accident Report NTSB/HAR-19/03 PB2019-101402 online:
<https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR1903.pdf>

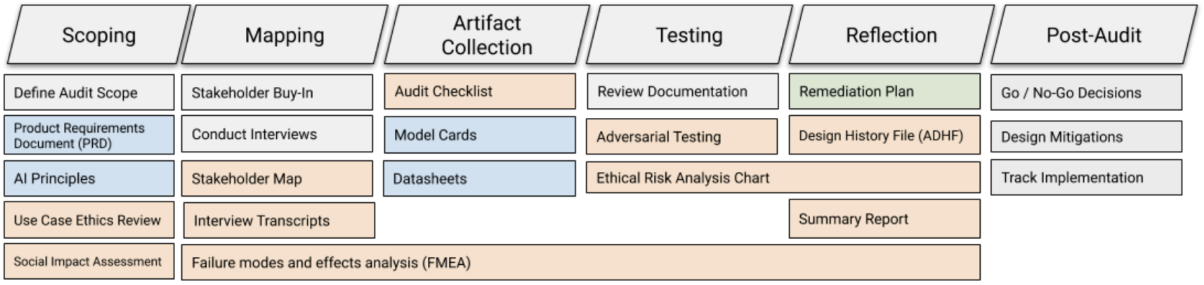
[81] S. Maurer, R. Erbach, I. Kraiem, S. Kuhnert, P. Grimm, E. Rukzio, "Designing a guardian angel: Giving an automated vehicle the possibility to override its driver"
Proceedings of the 10th international conference on automotive user interfaces and interactive vehicular applications, ACM (2018), pp. 341-350.

[82] Lilian Edwards & Michael Veale, "Slave To The Algorithm? Why A 'right To An Explanation' Is Probably Not The Remedy You Are Looking For" online:
<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2972855>. See also: Bryan Casey, Ashkon Farhangi, & Roland Vogl, "Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise", 34 Berkeley Technology Law Journal 143 (2019) online:
<https://btlj.org/data/articles2019/34_1/04_Casey_Web.pdf>.

[83] Deborah Raji et al. "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing" online: <https://arxiv.org/abs/2001.00973>.

[84] *GDPR Guidelines supra* note 64 at 32.

attributes).[85] Audits can be run internally or by external parties. External audits of AI systems that uncovered bias have prompted the developers to release new versions that showed better performance for protected groups.[86] An example audit process is below,[87] which notably includes several types of explanation including model cards, datasheets, and failure modes and effects analysis. Since 2019, American lawmakers have been considering legislation that would require assessments of high-risk automated-decision systems that would include "a detailed description of the automated decision system, its design, its training, data, and its purpose".[88]



European lawmakers have also suggested that data controllers explore "certification mechanisms for processing operations."[89] While some smaller countries, such as Malta,[90] are developing certification frameworks, there is no consensus even within Europe on what certification of AI systems should entail.[91] Non-governmental efforts, such as the initiative by the World Economic Forum, AI Global and the Schwartz Reisman Institute to create "a globally recognized certification mark for the responsible and trusted use of AI systems"[92] are still nascent. In the absence of regulatory consensus, scholars have recommended looking to frameworks that are already used in other safety-critical contexts, such as the Claims-Arguments-Evidence framework.[93] Progress has been made on techniques to generate evidence for claims about AI systems, such generating test inputs to identify corner

[85] Anya Prince & Daniel Schwarcz, "Proxy Discrimination in the Age of Artificial Intelligence and Big Data" 105 Iowa Law Review 1257 (2020)
[86] Raji, I & Buolamwini, J. (2019). "Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products." Conference on Artificial Intelligence, Ethics, and Society.
[87] Diagram from: Inioluwa Deborah Raji et al (2020) "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing" Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency online: <https://doi.org/10.1145/3351095.3372873>
[88] Algorithmic Accountability Act of 2019 2.A, online: <https://www.congress.gov/bill/116th-congress/house-bill/223>.
[89] *GDPR Guidelines supra* note 64 at 32.
[90] Malta Digital Innovation Authority, "AI ITA Guidelines" online: <https://mdia.gov.mt/wp-content/uploads/2019/10/AI-ITA-Guidelines-03OCT19.pdf>.
[91] European Commission. "On Artificial Intelligence - A European approach to excellence and trust" online: <thttps://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf>
[92] "Independent Certification Working Group Launched for Advancing Ethical and Responsible AI" online: <https://www.newswire.com/news/independent-certification-working-group-launched-for-advancing-ethical-21267921>.
[93] In this framework claims are typically statements about a property of the system such as "this system is safe" or ""this system was developed in accordance with our organization's ethical principles" with subclaims such as "this system is has a certain degree of privacy protection or non-discrimination". Evidence justifies a claim. Sources of evidence can include the design, the development process, prior experience, testing, or formal analysis. Arguments are then used to explicitly link evidence to a claim. See Miles Brundage et al. *supra* note 11.

cases[94] or identifying situations in which incorrect (but potentially plausible) weights could be loaded,[95] which this chapter would categorize as types of internal and external explanations respectively. Therefore financial institutions should consider how certification and algorithmic audits might both require some kinds of explanations to be generated.

2.5. Appropriately calibrating trust to manage risk

Apart from legal risks, the use of AI systems may create operational risks that can be managed using explanations. There are two ways in which people may inappropriately rely on automated systems.[96] Disuse refers to situations where a person fails to use automation that could have enhanced performance. For instance, an operator might turn off or ignore an automated warning system after it generated many false alarms. A lack of explainability can, and has, led to disuse of otherwise valuable AI tools within financial institutions. For instance: "[o]ne bank worked for months on a machine-learning product-recommendation engine designed to help relationship managers cross-sell. But because the managers could not explain the rationale behind the model's recommendations, they disregarded them."[97] In this and similar situations, providing explanations could reduce the risk of disuse.

Misuse refers to over-reliance on automation, either when the human operator fails to notice an error or follows an incorrect recommendation. An example of this is a global bank that misused a risk-hedging tool and only adjusted control parameters rather than change its investment strategy. As a result, it passed its value-at-risk limits for nearly a week and accrued a loss totaling in the billions of dollars.[98] Research shows that people are naturally biased toward over-reliance, and are more likely to believe answers that originated from a machine.[99] A simple approach to avoiding this is using explanation techniques that report on the model's uncertainty in its prediction or decision to encourage users to ignore or override the model when it is unlikely to perform well. Organizations can also correct for over-reliance bias by using explanations to appropriately calibrate people's trust with AI systems' actual trustworthiness — which in some cases might imply lowering their reliance on the system based on known limitations. For instance, a financial institution might want its underwriters to understand how susceptible a model is to manipulations of the input values

---

[94] Youcheng Sun et al. "Concolic Testing for Deep Neural Networks" online: <http://www.kroening.com/papers/ase2018.pdf>.
[95] Robin Bloomfield, Heidy Khlaaf, Philippa Ryan Conmy, and Gareth Fletcher, "Disruptive Innovations and Disruptive Assurance: Assuring Machine Learning and Autonomy" online: <https://www.heidyk.com/publications/08812789.pdf> at 88.
[96] Bronwyn French, Andreas Duenser, Andrew Heathcote, "Trust in Automation A Literature Review" Retrieved from https://pdfs.semanticscholar.org/92f0/7d3d1356307decb6e97382ad884d0f62668d.pdf.
[97] McKinsey. "Derisking machine learning and artificial intelligence" <https://www.mckinsey.com/business-functions/risk/our-insights/derisking-machine-learning-and-artificial-intelligence>.
[98] McKinsey, "The evolution of model risk management" online: <https://www.mckinsey.com/business-functions/risk/our-insights/the-evolution-of-model-risk-management>,
[99] Cummings M. "Automation bias in intelligent time critical decision support systems." In AIAA 1st Intelligent Systems Technical Conf., 2004. p. 6313.

that do not correspond to any changes in a borrower's likelihood to repay a loan.[100] This might be done by credit-seekers requesting periodic increases to credit lines just to decrease their debt ratio.[101] Explanations of this susceptibility would help lenders avoid accidentally taking on excess credit risk.

Models may also learn hidden spurious features that make them less reliable when applied in new settings.[102] If this happens, external explanations might be needed to explain why, for instance, stress-testing or back-testing failed to catch the issue. Without explicitly mandating the use of explanations to calibrate trust in this way, key regulators, such as the U.S. Federal Reserve[103] and U.K. Financial Stability Board[104] have highlighted the links between a lack of interpretability and risk in the financial sector at the macro-level and the individual firm level. For instance, U.S. guidance requires that decision makers understand the limitations of a model to avoid misuse. Ongoing monitoring of an AI system is also a key part of model risk management. Monitoring should include capabilities such as outlier and drift detection to alert operators when the model's predictions become unreliable because the relationships between inputs and outputs have diverged from the distributions on which the algorithm had been trained.[105] Although they are rarely framed explicitly in this way, drift detection is a type of external explanation: these techniques focus on identifying divergences between what a model has learned and the actual state of relationships in the world — that is, situations where a model is more likely to have "mistaken beliefs" in the sense of section 1.2.

2.6. Sustaining high performance in semi-automated contexts

Some AI applications, such as high-frequency trading, operate at a time scale that makes human involvement in decision-making impractical. However, most applications of AI in financial services are semi-automated in the sense that humans still play an important role either in making decisions (often referred to as "human in the loop" processing) or supervising algorithms (referred to as "human over the loop" processing). Combining human

[100] Explanation techniques can measure how much potential a feature has to cause misclassification, and thus how exposed the model is to threats like adversarial examples. Adversarial examples are minimally different inputs that look benign but are designed to produce a misclassification or misprediction by the algorithm. See Alexander Hartl, Maximilian Bachl, Joachim Fabini, Tanja Zseby "Explainability and Adversarial Robustness for RNNs" online: <https://arxiv.org/abs/1912.09855>/

[101] The example is from Zachary C. Lipton, *supra* note 50 at 3.

[102] For example, computer-aided fracture diagnosis algorithms have shown to be inexplicably using patient traits (such as demographics) and process variables (such as the brand of X-ray machine) that are not obviously present in images, which makes it much more difficult for clinicians to know how to interpret the model's predictions, and ultimately could reduce adoption. M.A. Badgeley, et al. "Deep learning predicts hip fracture using confounding patient and healthcare variables." npj Digit. Med. 2, 31 (2019).

[103] Federal Reserve, "SR 11-7: Guidance on Model Risk Management" (2011) online: <https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf>

[104] "The lack of interpretability or 'auditability' of AI and machine learning methods has the potential to contribute to macro-level risk if not appropriately supervised by microprudential supervisors.... Notably, many AI and machine learning developed models are being 'trained' in a period of low volatility. As such, the models may not suggest optimal actions in a significant economic downturn or in a financial crisis, or the models may not suggest appropriate management of long-term risks." FSB (2017). "Artificial Intelligence and Machine Learning in Financial Services–Market Developments and Financial Stability Implication." Technical report, Financial Stability Board.

[105] Janis Klaise et al, "Monitoring and explainability of models in production" online: <https://arxiv.org/pdf/2007.06299.pdf>.

and machine intelligence can help financial institutions improve performance. For instance, robo-analysts are computer programs that conduct automated research analysis with some support from human analysts. Research has found that robo-analysts produce a more balanced distribution of buy, hold, and sell recommendations, and that "portfolios formed based on the buy recommendations of robo-analysts appear to outperform those of human analysts."[106] The performance of a joint human-AI team increases when users have explanations, like the artificial agent's reasoning, rationale, or level of uncertainty.[107] Explanations can increase the rate at which users accept the advice given, or action chosen, by an automated system.[108] Facilitating appropriate levels of trust is central to maximising the performance and safety of human-AI teams,[109] and explanations can maintain trust levels when the system makes small errors that do not affect performance.[110] Perhaps more importantly, explanations can help signal performance issues that should decrease operator's trust and reliance on the system.

2.7. Summary of explanation needs

The table below summarizes the most likely pairings between the explainee persona, the motivation for providing explanations, and level of type of explainability required.

Legend: T: Transparency; I: Interpretability; C: Comprehensibility (see section 1.4 for definitions)

	Individual	**Institutional**							
	Decision subject/ user	**AI Developer / Designer**	**Business user**	**Business manager**	**1st line model checker**	**2nd line model checker**	**External auditor**	**Conduct regulator**	**Prudential regulator**
Data protection rights	**C**						**I**	**I**	**T**
Conduct basis requirements		**T**		**C**		**C**	**C**	**C**	
Negligent principal claims					**I**	**I**	**I**		
Algorithmic							**T**	**T**	**T**

[106] Coleman, Braiden and Merkley, Kenneth J. and Pacelli, Joseph. "Man versus Machine: A Comparison of Robo-Analyst and Traditional Research Analyst Investment Recommendations." (February 2020). Accessed at: http://dx.doi.org/10.2139/ssrn.3514879

[107] J. E. Mercado, M. A. Rupp, J. Y. Chen, M. J. Barnes, D. Barber, K. Procci, "Intelligent agent transparency in human–agent teaming for Multi-UxV management" Human Factors 58 (3) (2016) 401–415

[108] J. M.Bindewald, C. F. Rusnock, & M. E. Miller, "Measuring human trust behavior in human-machine teams." in D. N. Cassenti (Ed.), *Advances in Human Factors in Simulation and Modelin*g (Vol. 591, pp. 47–58). Cham: Springer International Publishing.

[109] Hoff, K. A., & Bashir, M. (2015). "Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors" The Journal of the Human Factors and Ergonomics Society, 57, 407–434. https://doi.org/10.1177/0018720814547570.

[110] Nora Balfe, Sarah Sharples, John R. Wilson "Understanding Is Key: An Analysis of Factors Pertaining to Trust in a Real-World Automation System". Balfe, Human factors vol. 60,4 (2018) online: <https://doi.org/10.1177/0018720818761256>.

auditing									
Calibrating trust to manage risk	**C**	**I**	**C**	**C**	**I**		**I**		**I**
Sustaining high joint human-AI performance	**C**	**C**	**C**	**C**					

Given this complex matrix, it is important for enterprises to think carefully about what explanation capabilities are required in order to put safe, trustworthy and high-performing AI systems into production. Of course, identifying requirements is only the first step. Organizations need to also consider how they are actually going to meet those needs by using different explainability techniques and strategies.

3. How to make explanations effective?

In aiming to make a system more interpretable or more comprehensible, there are several features of explanations that are beneficial. The review of desirable features below draws on cognitive science, sociology, and human factors studies, not just computer science. This is based on the underlying assumption that the policy-makers, regulators, and risk managers who have created the direct and indirect obligations for explanations want these explanations to be both true and useful. To be useful, AI explanations should correspond to the implicit "rules" of how humans explain things to each other,[111] including both people's ordinary social expectations, and their predictable cognitive biases.[112]

3.1. Selective

People prefer to receive explanations that cite fewer causes,[113] so explanations are better if they are tailored to provide a specific and limited amount of information based on each explainee's expectations. For instance, a user that receives a buy recommendation for a security might wonder why the algorithm suggested they buy rather than sell, or buy this stock rather than a different one, or buy a stock rather than a bond. Each of these (sell, other stock, other bond) are different foils[114] — that is, a specific alternative ("buy") to the actual prediction ("sell") that an explainee considers as a reasonable alternative. Usually the foil that the explainee has in mind is based on what they consider abnormal or unexpected from their own point of view and background knowledge.[115] When a consumer is harmed by some algorithmic decision, the consumer, business manager and regulator might have different perspectives on what needs to be explained. Perhaps the consumer wonders why a stock

[111] *Miller supra* note 10.
[112] *Ibid* at 32.
[113] S. J. Read, A. Marcus-Newhall, "Explanatory coherence in social explanations: A parallel distributed processing account" Journal of Personality and Social Psychology 65 (3) (1993) 429. Referenced in *Miller* supra note 10.
[114] The term is from *Miller* supra note 10.
[115] D. J. Hilton, "Mental models and causal explanation: Judgements of probable cause and explanatory relevance" Thinking & Reasoning 2 (4) (1996) 273–308. Referenced in *Miller supra* note 10.

went down rather than up, an analyst might wonder why the volatility was higher than expected, and a regulator might be interested in whether the process was compliant or not. On the other hand it may be more useful to specifically explain what would need to change ("even if you trebled your salary, you would still not get the loan") rather than a bare cause ("your salary was insufficiently high), but it is important that these counterfactual changes remain plausible.[116]

3.2. Robust

An explanation is robust to the extent that the causes it identifies remain constant as conditions change. Some interpretability methods do not produce very stable explanations. For instance, data processing steps that should be irrelevant, such as shifting the mean of a distribution, have been found to produce significant changes in the outputs of some saliency maps.[117] As discussed above, LIME identifies a small set of linear features that do a good job of approximating the local behavior of a model — that is, the technique identifies the features that matter for a particular input. By design, this explanation simplifies a more complex model, but this means that input-output pairs that seem similar to a human observer may have significantly different explanations (i.e. lists of features that matter) depending on the model's higher-dimensional geometry. In a financial context this could mean that an AI model used to value a house might explain that one house's valuation was lowered because it was an older home and local zoning was mostly for large lots, while a very similar house's valuation was increased by these same factors.[118] Both explanations could be true because of complex interactions (perhaps one was in an industrial zone while another was in a traditional daily neighborhood), but an explainee will experience these differences as a lack of robustness.[119] Explanations that vary considerably and are inconsistent with each other for neighboring inputs may be less useful in building trust. In contrast, explanations that operate at a higher level of abstraction (such as concepts rather than individual pixels) are more likely to be robust to small changes[120] and comprehensible to explainees who have similar mental concepts.

3.3. Complete

While robustness asks whether an explanation is appropriately agnostic to irrelevant changes, completeness asks whether how well an explanation of a point or portion of input space covers the whole underlying predictive model. Generating more complete explanations

---

[116] Eoin M. Kenny & Mark T. Keane, "On Generating Plausible Counterfactual and Semi-Factual Explanations for Deep Learning" online: <https://arxiv.org/pdf/2009.06399.pdf> at 6.
[117] Pieter-Jan Kindermans et al ¨The (un) reliability of saliency methods." arXiv preprint arXiv:1711.00867, 2017. See also Julius Adebayo et al. "Sanity Checks for Saliency Maps". In: arXiv (Oct. 2018). arXiv: 1810.03292. Online: < http://arxiv.org/abs/1810.03292>.
[118] David Alvarez-Melis & Tommi S. Jaakkola "On the Robustness of Interpretability Methods" online: <https://arxiv.org/pdf/1806.08049.pdf>
[119] See for example, Philippe Bracke, Anupam Datta, Carsten Jung, and Shayak Sen. "Machine learning explainability in finance: an application to default risk analysis" Staff Working Paper No. 816 The authors measured feature influences by intervening on inputs and found the expected drivers of mortgage defaults such as the loan-to-value ratio and current interest rate, "however, given the non-linearity of ML model, explanations vary significantly for different groups of loans."
[120] See Been Kim *supra* note 52.

aligns with people's general preferences for receiving explanations based on causes that are necessary, not just sufficient.[121]

The amount of content in an explanation can vary significantly and be quite large if, for instance, the explanations are targeted at an expert user inside of an organization rather than an external customer. One experiment found that performance and trust calibration improved steadily as the richness of the explanations expanded. The explanation content increased from just the system's current plans, to also include the system's reasoning process (rationale, capabilities, limitations, and trade-offs), and finally to encompass the system's projections of future outcomes (predicted consequences and the likelihood of a plan's success or failure).[122] However, the compositionality of explanation raises an important difficulty related to completeness. That is, if a business outcome is the product of two or more models, the explanations of those component models may need to be computed and combined to produce a complete explanation for the overall decision.

3.4. Sound

In some situations, predictions and explanations are generated by the same algorithm, but in other situations a secondary (surrogate) model may be used to explain a primary model. For instance a decision tree might be generated to approximate the whole of a more complex network neural network[123] or a local linear model could approximate a region of it. A real-world example comes from a team at the U.K. Financial Conduct Authority and Bank of England who developed a model to predict which borrowers would fall behind on mortgage payments. They then used explainability techniques to identify the main factors driving its predictions and built a simplified model based on these factors to explain why some mortgages are classified as high-risk and others as low-risk that would be more comprehensible to non-specialists.[124] In any dual-model setup like this the predictive and explanatory models may diverge. The amount of divergence can be measured,[125] and explanatory models that diverge less are more sound — that is, they are more truthful in regard to the original processing model.

Similarly, there exist "dually adversarial attacks" which generate inputs that manipulate a model's prediction while keeping its explanation intact.[126] Explanations that can be decoupled from data and predictions in these ways are less sound. Studies have found that

---

[121] P. Lipton, "Contrastive explanation", Royal Institute of Philosophy Supplement 27 (1990) 247–266. Referenced in *Miller supra* note 10.
[122] J. Y. C Chen, et al. "Situation awareness–based agent transparency" (2014). (No. ARL-TR-6905). Aberdeen Proving Ground, MD: U.S. Army Research Laboratory.
[123] Zilke J.R., Loza Mencía E., Janssen F. (2016) "DeepRED – Rule Extraction from Deep Neural Networks." In: Calders T., Ceci M., Malerba D. (eds) *Discovery Scienc*e. DS 2016. Lecture Notes in Computer Science, vol 9956. Springer, Cham.
[124] K. Croxson, P. Bracke, & C. Jung, "Explaining Why the Computer Says 'no'." (2019). London, UK: FCA-Insight.
[125] Divergence can, for example, be calculated based on the average rank correlation between the two. See Peter Flach & Kacper Sokol, "Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches" online: <https://arxiv.org/abs/1912.05100>.
[126] See Xinyang Zhang et al., "Interpretable Deep Learning under Fire" https://arxiv.org/pdf/1812.00891.pdf

users did not trust and were less likely to attend to less sound explanations.[127] Soundness is especially important when the explainee's goal is to identify and fix bugs or errors in the model to ensure they are building an accurate internal representation of what the predictive algorithm is doing.[128]

3.5. Actionable

Designers of AI interfaces have highlighted the importance of making explanations interactive because performance scores (like model accuracy) are rarely enough to produce the desired levels of trust and adoption.[129] However users will tend to ignore explanations unless there is a clear benefit to paying attention to them.[130] One of the most important potential benefits of an explanation is that it enables the explainee to intervene in some way, by taking some action after receiving the explanation such as challenging the decision or requesting human intervention (as envisaged in GDPR and related laws) or improving the model itself. Therefore actionable explanations are most likely to produce good trust calibration and adoption.

To be actionable, a reason for a decision needs to be changeable[131] for the user. Within the features used by a model there may be a spectrum of variables from the immutable (such as changing an adult's height) to impractical (such as acquiring a PhD to qualify for a loan) to possible (such as increasing one's income) to feasible (such as consolidating debt). The feasibility of actions can sometimes be learned directly from data.[132] However, in general actionability needs to be assessed against insights from the world outside of the model to reflect socio-economic and physical realities of decision subjects, or provide users themselves with mechanisms for determining what counts as actionable.

4. Challenges

There are many challenges to producing explanations that are selective, robust, complete, sound, and actionable for the many different types of AI models, data domains, and user personas. A full review is out of scope for this chapter, but a few points are worth highlighting.

---

[127] Kulesza, T., Stumpf, S., Burnett, M. M., and Yang, S. "Too much, too little, or just right? Ways explanations impact end users' mental models." Proceedings of the 2013 IEEE Symposium on Visual Languages and Human-Centric Computing (2013), 3–10.
[128] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. "Principles of Explanatory Debugging to Personalize Interactive Machine Learning." In Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15). Association for Computing Machinery, New York, NY, USA, 126–137.
[129] Vera Liao, Daniel Gruen & Sarah Miller, *supra* note 9.
[130] Kulesza et al., *supra* note 129.
[131] This could imply, but does not require, a causal connection between the key features and the prediction. Learning causal relationships is an active research area, see for example Stephen L. Morgan & Christopher Winship *supra* note 33.
[132] For instance since there are likely to be no teenagers with PhDs in a dataset, an explanation that a teenage applicant would have been approved for a loan if they had a tertiary degree could be inferred to not be actionable. See Ramaravind Kommiya Mothilal, Amit Sharma, Chenhao Tan, "Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations" online: <https://arxiv.org/abs/1905.07697>.

4.1. The challenge of siloed expertise

Since explanations are fundamentally a social undertaking involving human agents, data scientists need to collaborate with designers, psychologists, and other experts outside of math and computer science in order to make better choices in system design. For example, incorporating the explainee's mental model as part of the explainability system to make it appropriately selective requires a model of human cognition that is at least fit for purpose.[133]

Some work has been done at the intersection of these fields such as testing users' preferences for different combinations of simplicity, completeness and soundness.[134] For instance, one study presented users with two formats of explanations when a model misclassified an input image: one format presented example inputs that had been classified correctly, the other format presented example images of the classes that the model believed were most similar to the input image. Users that saw the first format rated the system's capability as higher, but users that saw the second format trusted the model more, likely because this format exposed more of the model's limitations.[135] This type of insight might be adapted by industry, but financial institutions still face many uncertainties in how best to implement academic research. For instance, metrics to measure the quality of explanations are still being developed,[136] which leaves developers without an independent, objective way to consider and compare different design choices.

4.2. The challenge of limited guidance from supervisory bodies

The subtle and sometimes highly technical relationships between causality, reasons, and explanations in algorithmic systems can create issues for regulatory bodies that are more familiar with human agents and more deterministic software. For instance, in 2019 the U.S. Department of Housing and Urban Development proposed a safe harbor rule that would protect an algorithm's creators from discrimination claims if it excluded protected attributes.[137] However, the presence or absence of variables (or close proxies) in a model that indicates a person's membership in a protected class does not indicate the presence or absence of a causal connection between a decision and class membership,[138] so this proposal might accidentally provide a liability shield to discriminatory algorithms based on a misunderstanding of what could be inferred from their inner logic. More recently, a 2021 U.S. Presidential Executive Order on advancing racial equality recognized that "[m]any

---

[133] One example of early work on user modelling was the EDGE system, which used dialogue questions to infer the knowledge that the user has about a phenomenon and their level of expertise to tailor explanations to a person. Cawsey (1991) A. Cawsey, Generating Interactive Explanations, in: AAAI, 86–91, 1991, cited in *Miller* supra note 9

[134] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, W.-K. Wong, "Too much, too little, or just right? Ways explanations impact end users' mental models" in: Visual Languages and HumanCentric Computing (VL/HCC), 2013 IEEE Symposium on, IEEE, 3–10, 2013.

[135] Carrie J. Cai et al. "The Effects of Example-Based Explanations in a Machine Learning Interface" 2019. Proceedings of the 24th International Conference on Intelligent User Interfaces.

[136] Andreas Holzinger, André Carrington & Heimo Müller "Measuring the Quality of Explanations: The System Causability Scale (SCS). Comparing Human and Machine Explanations" online: <https://arxiv.org/abs/1912.09024>

[137] Alice Xiang "On the Legal Compatibility of Fairness Definitions" online: <https://arxiv.org/pdf/1912.00761.pdf>

[138] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva *supra* note 43.

Federal datasets are not disaggregated by race, ethnicity, gender, disability, income, veteran status, or other key demographic variables. This lack of data has cascading effects and impedes efforts to measure and advance equity."[139] The order thus recognizes that unawareness of sensitive attributes can be counterproductive to advancing goals such as anti-discrimination.

While this type of order suggests a growing understanding of some of the intricacies of explainable AI within government, the details of explainability regulation are still often underdetermined. For example, GDPR seems to set a minimal threshold for the content of explanations but not a maximum. The Guidelines on what constitutes "meaningful information about the logic involved" in processing clarify that the controller does not necessarily need to provide a complex explanation of the algorithms used, but should provide information that is "sufficiently comprehensive for the data subject to understand the reasons for the decision."[140] What constitutes "sufficiently comprehensive" is likely context-specific, which leaves financial institutions with a gap in guidance.

Regulators like Singapore's Monetary Authority are developing detailed guidance and case studies on explainability in financial services.[141] However it will be challenging for any regulator to have enough detail and specificity to cover all use cases. For example, should a robo-advisor's recommendations be accompanied by explanations of why a specific security was selected rather than other securities in the same asset class only? Should alternative securities be named? How many of these alternatives are useful? Currently each company must discover and define for themselves how to select the most relevant format and content of information while also not overwhelming users. These types of gaps are understandable given the rapid development of AI technology, but the lack of sector-specific guidance on what counts as sufficient explanations is a challenge for financial institutions today that cannot be overcome purely through internal efforts.

4.3 The challenge of explaining machine-learned reasoning

Unlike humans, AI systems draw on a limited experience from the training set to establish their logic. As such, the internal reasoning of a model might produce a high accuracy outcome but do so for reasons that seem spurious to a human evaluator. For instance, an image classification algorithm might find that the presence or absence or snow in the background is the best predictor of whether an image contains a wolf-like dog (such as a husky) or a wolf.[142] This might be predictive in the training set and be a true explanation of the model's reasoning, but nonetheless fail to convince a human explainee — as it would, correctly, suggest that the model has failed to learn a generalizable classification function. This type of disconnect risks reducing the trust of end-users such as a consumer (rather than a data scientist or engineer) who receive the explanations.

---

[139] White House, "Executive Order On Advancing Racial Equity and Support for Underserved Communities Through the Federal Government" (2021) section 9.
[140] GDPR Guidelines *supra* note 64 at 25.
[141] Veritas Consortium "FEAT Fairness Principles Assessment Methodology" online: <https://www.mas.gov.sg/-/media/MAS/News/Media-Releases/2021/Veritas-Document-1-FEAT-Fairness-Principles-Assessment-Methodology.pdf>
[142] The example is from Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin *supra* note 19.

Explanation techniques may generate highly technical and detailed outputs that may be legible to practitioners but not for ordinary users or even experts from other domains They may also be legible to experts but draw upon features of the data the expert disagrees with. An example would be an insurance application system that rejects an applicant based on a complex combination of features while a human expert rejects the same applicant based on a single feature (such as a credit score). While both the expert and the system agree on the rejection, the underlying motivations are sufficiently distinct to cause distrust. As a result, some explanations may require an additional layer of translation and interpretation to be useful, which can introduce its own challenges of balancing accuracy and completeness.

5. Recommendations to financial institutions on creating explainable AI

5.1. Catalogue your explicit and implicit explanation obligations

For any AI system, there are at least six potential motivations for providing explanations: meeting obligations to individual rights-holders; justifying conduct (such as why a specific product was recommended); rebutting claims that the organization was negligent in training or monitoring the AI system; enabling institutional safeguards such as algorithmic audits; managing risk by helping users and other stakeholders appropriately calibrate their trust in the system; and providing the information needed by human collaborators to sustain high performance in semi-automated settings.

Each explanation requirement may have a distinct motivation, ranging from assuring the safety and soundness of the financial system, to providing general consumer protection, to avoiding disparate impact on individuals from specific groups that faced historical discirimination. Rather than jumping straight into picking among the many available AI techniques, financial institutions should understand the underlying motivations of the related legislation, standard, guidance or internal policy and whether it is best served by an internal, external, or normative explanation. Designers, data scientists, engineers, and product managers (referred to below collectively as “AI builders”) should identify and document explanation requirements early, ideally in the conceptual design phase for new AI systems.

5.2. Bring your developers, lawyers, risk managers and designers together for solutioning

Once the various explanation needs are understood, it is important to bring together the many different kinds of expertise required to build safe and effective socio-technical systems. Finding the right solution requires experts that can ask and answer a diverse set of questions, including human factors considerations (e.g., “what information will the user be able to absorb and act on?”), operational risks (e.g., “what data or vulnerability do we risk exposing through explanations?”), legal strategies (e.g., “will having or not having certain information help if an adverse event occurs?”), and technical theorizing (“what assumptions is the explanation technique making, and are these true in this case?”).

Effective explanations should ideally be integrated early into the system design rather than be added as an extra feature late in development. The need to generate, store, and reproduce explanations and enable users to interact with them should be influenced by the constraints of the production system (such as speed, memory and security), and should in turn impact choices about model and system architecture. If explainability requirements are not well defined, this can create substantial technical debt for developers charged with retrofitting explainability modules into complex systems. The multiplicity of related algorithms and software that may need to be implemented and connected for the system to work and generate explanations of its own working can create significant problems of compatibility and scalability due to the relative immaturity of this software from a system engineering perspective. This challenge is typical of software being built in any new and emergent deep tech field and strong engineering approach and experience developers are needed to handle the associated risk. The risk can also be reduced by avoiding the temptation to incorporate the very latest advances from academia, and instead using both AI frameworks and explanability algorithms that have been more fully tested in practical applications.[143]

5.3. Develop competencies for both individual- and institutional-level explanations

In designing explanations of an AI system, organizations with strong user-centric design practices may be likely to focus on individual explanations for consumers. Because decision-subject explanations are important and mandated by some legislation, having reusable capabilities that can be deployed across models is important for efficiently scaling the use of AI. However it is equally important to build capabilities to produce the types of explanations envisaged by institutional-level safeguards of automated decisions-making. For instance, model cards and other logging or traceability-style explanations are highly useful for algorithmic audits, and tools that let developers and auditors scientifically probe a model (such as by systematically recording the effects of small changes in puts or training) are useful for debugging and making AI systems more robust and resilient.

5.4. Embed explanation competencies into the broader model risk management and governance framework

AI teams focused on innovation in financial institutions are sometimes separated from business groups and even other analytics or data-science teams that are building more traditional models fully within financial institutions' model governance framework. While this might appear to accelerate innovation, it risks limiting AI to internal labs or proof-of-concepts if the models that are developed are prohibited or severely delayed by the model risk management process.

To avoid this, AI builders should be aware of the type of validation and verification that is required for deployment to production, and the validation and verification teams should be equipped with explainability tools to help with their work. Risks uncovered through explainability work (such as how well the model is likely to perform through the business

[143] See A. Lavin, et al. *supra* note 36 for a framework that assesses the maturity of machine learning systems.

cycle, or the disparate impact on specific groups due to specific choices of business KPI) should be integrated with risk reporting.

5.5. Measure and manage the impact of providing explanations

More comprehensive risk reporting is one positive impact of increasing model explainability but it is far from the only one. Because building the capacity for explainability requires investment, it is also important to track and quantify the returns these investments are generating beyond compliance. This is an area where AI practitioners from industry can take the lead in filling a gap in academic research on testing explainability with real-world users and setting. A/B testing the impact of different explainability designs (such as how much or what type of information is provided), measuring developers' usage of explainability-based debugging tools, and instrumenting systems to identify spurious learned relationships are a few examples of how the impact of explanations can be measured. From there, it is a relatively short step to efficiently managing the resources that are put toward explainability. Anecdotally, we believe many financial institutions are under-investing in explainability capabilities compared to the potential benefits, and more measurement of explanations' impact will lead to more investment. In turn, investment in making AI more explainable will help financial institutions create more value and better manage risk.

Conclusion

Financial institutions have an acute need for AI systems that are explainable to ensure they are meeting their obligations to customers, shareholders and regulators. Consumers' right to have individual algorithmic decisions explained is only one source of these obligations; institutional motivations, such as effective legal and operational risk management and maximizing the performance of human-AI teams, are equally important reasons to want explanations. What different stakeholders want from explanations can range from understanding the internal logic of a model, to how a system (including the data generation and consumption processes) is learning patterns that fail to reflect reality, to whether a decision or process can be ethically justified. To achieve these varied jobs-to-be-done by explanations, financial institutions will need to combine academic work on explainability techniques with the contributions of designers, developers, engineers, and legal and business subject-matter experts. Doing so will help take advantage of the enormous potential of using algorithms that can learn from data rather than human instruction, without alienating the people who should be at the centre of these socio-technical systems.