



The Under-appreciated Regulatory Challenges posed by Algorithms in FinTech

Understanding interactions among users, firms, algorithm decision systems & regulators

Berlin, Fall 2021

Sahil Deo

Dissertation submitted to the Hertie School in partial fulfilment of the requirements for the degree of Doctor of Philosophy (PhD) in the Doctoral Programme in Governance Berlin, 2021

Advisors

First Advisor

Prof. Mark Hallerberg, PhD

Hertie School

Second Advisor

Prof. Kalpana Shankar, PhD

University College Dublin

Third Advisor

Prof. Sudhir Krishnaswamy, PhD

National Law School of India University

Summary

The rise of automated decision-making systems has been extraordinary in the last few years, both in terms of scale and scope of operation. This is especially true in the field of fintech, where robo financial advisors have gained prominence with their claim to “democratise” finance, with their low operating costs, multi-tasking abilities and potential for mass adoption. This series of three research papers is focused on demystifying algorithmic explainability in the field of fintech, by diving deep into both theoretical and practical aspects of the phenomenon, and contextualising the discussion to India.

The first paper explores the trade-off that emerges between performance and explainability for robo financial advisors, through a detailed review of available literature that allowed comparisons between relevant processes adopted around the world, and interviews with various stakeholders within India to understand the evolving domestic situation. The paper finds that it is not quite a question of *if* ADS will play a significant role in India’s financial services sector but more a question of *when* that will happen. The second paper operationalises algorithmic explainability in the particular context of risk profiling done by robo financial advisory applications. Here, an approach towards developing a ‘RegTech’ tool is outlined, which can explain the robo advisor’s decision-making, using machine learning models to recognise and reconstruct different levels of explanations. Finally, the third paper evaluates the effectiveness of user-centric explanations in conveying the decision-making logic of complex algorithmic systems in fintech. The paper demonstrates the usefulness of such explanations from the perspectives of both novice and seasoned investors, and goes on to differentiate between white- and black-box explanations.

In sum, this three-paper series, using a range of tools and approaches, examines in detail the under-appreciated regulatory and operational challenges that emerge during the use of algorithms in the field of fintech, and explores ways to resolve them. While the first paper looks at the present and future of AI in Indian fintech, the second paper develops a tool to explain a robo advisor’s decision-making, and the third finds factors that determine users’ comprehension and

confidence in such systems. The results of these three approaches in the three papers become vital when seen in the context of the rapid rise of artificial intelligence across products, services and industries globally. If humans are to work alongside machines in this changing world, explainability becomes an important aspect to consider for companies, regulators and users alike, in order for humans to trust the algorithmic systems in place and improve outcomes — in this case, financial outcomes — for all.

Acknowledgements

First and foremost I think my mother, Usha, & wife, Mrunmayee, without whom this "project" would have never seen light. Alongside them, Prof. Hallerberg has been a firm pillar of support and a guiding light. I thank my advisors, especially Prof. Shankar, and other academic stalwarts who have guided me along in this journey.

I so wish Shri Vasant Deo was here to see this take shape. Without the blessings of Gurudev, none of this was possible.

Table of Contents

<i>Advisors Page</i>	<i>ii</i>
<i>Summary</i>	<i>iii</i>
<i>Acknowledgements</i>	<i>v</i>
<i>Table of Contents</i>	<i>vi</i>
<i>List of tables</i>	<i>ix</i>
<i>List of figures</i>	<i>x</i>
<i>List of abbreviations</i>	<i>xii</i>
<i>Introduction</i>	1
<i>Trade-off between algorithmic performance and explainability for robo-financial advisors in the Indian context</i>	11
Abstract	11
Introduction	12
Section I: Conceptual background	12
Section II: Algorithmic governance and the rising demand for explainability	15
Section III: Algorithmic Governance — an international comparison	21
Section IV: Review of algorithmic regulation in India	40
Section V: Contextualising algorithmic regulation to fintech & robo financial advisors	48
Section VI: Situating the right to explainability in the Indian context	61
Section VII: Adjusting the FATE framework to the Indian context	65
Section VIII: Empirical Research	69
Research Methodology	74
Hypotheses	76
Interviews	77
Ethics	78
Sampling	78
Sampling frame & sample matrix	79
Thematic analysis	79
Potential interviewees	81
Analysis of interviews	82
Pervasion of ADS in the financial services sector	82
Maturity of the responsible AI movement in India	85
Appropriate mode of regulation	87
Unique constraints operating in the Indian context	89
Trade-off between algorithmic explainability and algorithmic performance	91
Conclusion	95
Appendix (A)	97
List of Stakeholders	97
Frameworks to assess algorithms	97
References	103
<i>Operationalising algorithmic explainability in the context of risk profiling done by robo financial advisory apps</i>	114
Abstract	114

Introduction	115
What are robo financial advisors?	116
Methodology	126
1. Generating the dataset for the study	127
2. Information that needs to be explained by the robo-advisory	129
3. Operationalising the explanations	131
Findings	135
Part 1- modelling the risk profiling decision.....	135
Part 2- Feature importance scores	140
Part 3- Relationships	143
Part 4- Local explanations.....	146
Conclusion	149
Discussion and way forward	150
Appendix	151
Appendix 1- Definitions and key terms	151
Appendix 2- Details of sample dataset generation that has been used for this study.....	152
Appendix 3- Explaining the machine learning models.....	160
References	166
<i>Algorithmic Explainability in Practice: Evaluating the effectiveness of explanations in the context of robo advisory apps</i>	171
Abstract	171
Introduction	172
Emergence of Robo Financial Advisors	174
Issues Surrounding Algorithmic Decision-Making and Trust Deficit.....	175
Addressing Trust Deficit: Algorithm-centric and User-centric Approaches	176
Explainable Artificial Intelligence: Review of Literature and Discussions	178
1. What is explainable AI?	179
2. Approaches to Algorithmic Explainability of Automated Decision Systems (ADS)	181
3. Measuring Effectiveness of Explanations through User Studies	186
Research Questions and Goals of our user study	188
Methodology & Experiment Design	190
1 Scope of our experiment - human centric XAI strategies.	190
2 Describing our Experiment	191
3 Replica Robo-Advisory System	192
4. Generating Explanations	196
White Box versus Black Box	214
Black Box	215
Experiment Design of User Study.....	225
Similar User Studies on XAI.....	226
Details on the survey design	227
Results	228
Participant demographics	229
User comprehension of explanations, recommendations	237

Effects of explanations on participant opinions or perceptions.....	240
Usability of explanations.....	243
Effects of explanations on users' trust in explanations, system recommendations	246
Difference in effect of explanations based on demographic backgrounds	249
Effect of complexity on user comprehension and usability (black box vs. white box)	251
Broader implications of findings for robo-advisory applications	252
Towards a standardised explanation strategy	253
Conclusion and Future Work	257
Appendix	258
What is XAI? What is the need for it?	258
Literature Review on Measuring Effectiveness of Explanations	262
Robo Advisor: User Risk Questionnaire	264
Robo financial advisors: going into the details.....	265
First Generation Systems	266
(1970 - 1985).....	266
Third generation (2015+).....	267
T-test results.....	275
References	278
Conclusion.....	290
Appendix.....	296
References for Introduction (pages 4-10)	297
References for Conclusion (pages 207-210).....	299

List of tables

Table 1. 1 Analysis and comparison of the approach adopted by various jurisdictions	32
Table 1. 2 Comparing the observations on prioritising explainability versus prioritising performance	93
Table 2. 1 Summary of results for the ML models and input equations	135
Table 3. 1 General trends across classes for each feature.....	206
Table 3. 2 Trends captured by model during mutual fund classification.....	210
Table 3. 3 Feature effects and behaviours captured by surrogate model of the user risk data	218
Table 3. 4 Risk trends captured by the surrogate model	222

List of figures

Figure 1. 1 Frameworks discussed in this paper to assess fairness of algorithms	16
Figure 1. 2 Key themes in the research methodology of this paper	75
Figure 1. 3 Summary of hypotheses	77
Figure 1. 4 Positive (right column) and unfavourable (left column) factors at play in the infiltration of ADS in financial services	84
Figure 1. 5 Reasoning offered for light-touch regulation	89
Figure 1. 6 Constraints cited for AI adoption in financial services in India	90
Figure 2. 1 Framework for explanation	131
Figure 2. 2 F1 scores and accuracy versus size of training data	139
Figure 2. 3 Overall feature impact on outcomes	141
Figure 2. 4 Class-wise feature importance plots	142
Figure 2. 5 Correlation between features and outcomes	144
Figure 2. 6 Relationship of age with output risk classes	144
Figure 2. 7 Relationship between age, dependents and low-risk output	145
Figure 2. 8 Features influencing predicted output class	146
Figure 2. 9 Influence of input features on all possible output classes	147
Figure 3. 1 Publications by year (1977-2017)	180
Figure 3. 2 Broad classification of XAI approaches	181
Figure 3. 3 Post-hoc explainability approaches	183
Figure 3. 4 Experiment design	191
Figure 3. 5 Abstract procedure for the task in the experiment	192
Figure 3. 6 Robo advisory system designed for the user study	193
Figure 3. 7 Data distribution of samples	198
Figure 3. 8 Model details for mutual fund risk profiling algorithm	200
Figure 3. 9 Model details for user risk profiling algorithm	200
Figure 3. 10 Impact of variables on risk category assignment	202
Figure 3. 11 Feature importance in mutual fund risk classification and recommendation	203
Figure 3. 12 Risk category explanations given to users	204
Figure 3. 13 Categorical variables used in user risk calculation	205
Figure 3. 14 Values assigned to variables	210
Figure 3. 15 Top decision boundaries shown to users	213
Figure 3. 16 Black box version: Model details for user risk	215
Figure 3. 17 Black box version: Model details for mutual fund risk determination	216
Figure 3. 18 Black box: Impact of variables on user risk category assignment	216
Figure 3. 19 Black box: Feature importance in mutual fund risk classification and recommendation	217
Figure 3. 20 Comparing users from black box and white box studies	232

Figure 3. 21 Demographic divisions.....	236
Figure 3. 22 Majority demographics of survey participants.....	237
Figure 3. 23 Comparing black box and white box explanations for comprehension and accuracy	240
Figure 3. 24 Difference between user response in white box and black box scenarios	242
Figure 3. 25 Participant preference of explanations.....	245
Figure 3. 26 Comparing white and black box explanations for user trust	248
Figure 3. 27 Demographic divisions in users	251

List of abbreviations

ADS – Algorithmic Decision-making Systems

AI – Artificial Intelligence

AI4SG – Artificial Intelligence for Social Good

APPs – Australian Privacy Principles

AUM – Assets under management

CDEI – Centre for Data Ethics and Innovation

COPRA – Children’s Online Privacy Protection Act

CSIRO – Commonwealth Scientific and Industrial Research Organisation

DPA – Data Protection Authority

DT – Decision Trees

EC – European Commission

EHR – Electronic Health Records

EMI – Equated monthly instalments

EU – European Union

FAT – Fairness, Accountability and Transparency

FAT/ML – Fairness, Accountability, and Transparency in Machine Learning

FATE – Fairness, Accountability, Transparency and Ethics

FTC – Federal Trade Commission

GDPR – General Data Protection Regulation

HCI – Human Computer Interface

ICO – Information Commissioner’s Office (UK)

IOSCO – International Organization of Securities Commissions

IRDAI – Insurance Regulatory and Development Authority of India

IT – Information Technology

JPC – Joint Parliamentary Committee

KNN – k-nearest neighbours

LR – Logistic Regression

MeitY – Ministry of Electronics and Information Technology

ML – Machine Learning

NB – Naïve Bayes

OECD – Organisation for Economic Co-operation and Development

P/B ratio – Price to book ratio

PDPs – Partial Dependence Plots

P/E ratio – Price to earnings ratio

PFRDA – Pension Fund Regulatory Development Authority

PIPEDA – Personal Information Protection and Electronic Documents Act

PS – Private sector (for interview purpose in Paper 1)

RAs – Robo financial advisors

RBI – Reserve Bank of India

RG – Regulators (for interview purpose in Paper 1)

SEBI – Securities and Exchange Board of India

SVM – Support Vector Machines

TA – Thematic analysis

UK – United Kingdom

UR – Users or researchers (for interview purpose in Paper 1)

US – United States

XAI – Explainable artificial intelligence

Introduction

“By far the greatest danger of artificial intelligence is that people conclude too early that they understand it,” observes Machine Intelligence Research Institute’s renowned artificial intelligence (AI) theorist, Eliezer Yudkowsky (Yudkowsky, 2008). He is referring to a commonly observable notion that all stakeholders, from developers to end-users involved in an AI lifecycle, come with preconceived notions on how the AI functions. The field of AI has a reputation for making huge promises and then failing to deliver on them. This often results in one of two cognitive biases: uncritical belief in the AI’s abilities or a complete lack of faith. Both are equally dangerous for impartial and equitable development and adoption of new technology. The future of AI and its potential to benefit society will be determined through transparent communication, accountable analyses, and unprejudiced critique of its functioning.

AI’s ability to process vast amounts of data is exceptional, facilitating both automation and personalisation of decision-making. There is a growing ubiquity of decision-making algorithms that affects our lives and the choices we make. This trend is visible in a host of domains cutting across the public and private sectors, such as loan approvals for fintech companies, hiring employees and identifying criminals by law enforcement agencies. These algorithmic decision-making systems (ADS) curate our internet and social media feed, trade in the stock market, assess risk in banking, fintech and insurance, diagnose health ailments, predict crime prevention, and a lot more. These algorithmic tools primarily rely on rich reserves of personal data about individuals. Such automation has meant that human decision making is now being progressively replaced by data fed algorithms.

Consequently, the quality and accuracy of data fed to these systems form the basis of the algorithm’s ability to analyse patterns to arrive at decisions. Several cases have come to light where algorithm-powered decisions have given rise to undesirable consequences. An automated hiring tool used by Amazon discriminated heavily against women applying for software development jobs because the machines learn from past data that has a disproportionate number

of men in software positions (Dastin, 2018). Software used for crime prediction in the United States showed a machine bias against African-Americans, exacerbating the systemic bias in the racial composition of prisons (ProPublica, 2016). Google's online advertising system displayed ads for high-income jobs to men much more often than it did to women (Datta, Tschantz, & Datta, 2015). Social media algorithms inadvertently promote extremist ideology (Costello, Hawdon, Ratliff, & Grantham, 2016) and affecting election results (Baer, 2019). Recently, researchers found that racial bias in the United States' health algorithms reduced the number of Black patients identified for extra care by more than half (Obermeyer, Powers, Vogeli, & Mullainathan, 2019) (Kari, 2019).

Therefore, contrary to the promise of unbiased and objective decision making, these examples point to a tendency of algorithms to unintentionally learn and reinforce undesired and non-obvious biases, thus creating a trust deficit. The complex decision-making logic of these algorithms is often difficult to follow, making them "black boxes". This arises mainly due to a lack of transparency and accountability in ADS. Due to the lack of adequate regulation, algorithms are not adequately tested for bias and are not subjected to external due diligence. The complexity and opacity in the algorithms decision-making process and the esoteric nature of programming denies those affected by it access to explore the rights-based concerns posed by algorithms. Decisions in the public sphere affect an individual's access to services and opportunities, and they need to be scrutinised.

For instance, with numerous decision-making algorithms, the financial services sector is one of the torchbearers for applications driven by artificial intelligence. On the security front, for at least a decade, banks proactively monitor and detect fraud, money laundering and other malpractices using AI. Since 2016, CitiBank uses AI-based monitoring for real-time risk management across banking and commerce (CitiBank, 2018). Additionally, AI is used in this sector for high-quality customer engagement through personalisation, virtual customer service and chatbots. For example, the HDFC Bank chatbot 'Eva' works with Google Assistant on millions of Android devices to solve customers' queries and provides them with

better services (HDFC Bank). Similarly, Axis Bank allows its customers to talk about their banking issues anytime, anywhere through a multi-lingual AI-powered bot called AXAA (AXIS bank, 2020). AI also helps improve processes in back-office operations through intelligent automation, and data analysis from bank and social media records have allowed for new ways to measure creditworthiness. In India, the Bank of Baroda is another public sector lender advancing banking services and reducing the cost of managing accounts through AI. With the power of AI, finance can be democratised through personalisation, scalability, expert advice and low-cost application.

To illustrate the power of AI, one can consider the robo financial advisory algorithm or RAs. Robo-advisory applications are online investment advisory algorithms that are automated and designed to recommend “the best plans for trading, investment, portfolio rebalancing, or tax saving, for each individual as per their requirements and preferences”. Robo-advisers are considered to be cheaper and more accessible when compared to human advisors, and hold the potential to democratise financial services by providing financial advice to sections of the population that are currently outside the formal banking system. This would help arrest the consumer-producer gap regarding the spatial and temporal dimensions prevalent among the current intermediaries in the financial system (banks, agents etc.). The ability of these applications to devise region and culture-specific investment strategies allows financial products to be made adaptable to local conditions. Robo-advisory applications have found their footing in the Indian financial services sector as well. Indian fintech companies adopt robo-advisory due to advantages such as low operating costs, ease of scaling, and minimisation of human error and fraud. While the adoption is currently at a nascent stage, it is likely to become widespread in the near future. The potential benefits of financial inclusion that robo-advisory could usher in make it particularly relevant and potent for the Indian market.

In the context of India, it is important to understand the various levels of complexity of the product landscape with respect to robo-advisory across various levels of complexity. Within the larger category of robo-advisory, some platforms

offer a digital interface that provides investors with an automated portfolio proposal with automatically selected funds or stocks. At the next level of complexity, there are robo-advisory services that are algorithm-driven, which offer automatic execution and portfolio rebalancing services based on investment strategies that have been planned previously. At the highest level of complexity are robo-advisory services that are fully intelligent systems that self-learn and are driven by economic theories without any significant human intervention. India currently has robo-advisories that are at the first two levels of complexity (Hon, 2019).

Although the overall number of individual investors has been increasing at a 10-year compound annual growth rate of 11 percent in India, access to wealth management services nevertheless remains limited. At the National Stock Exchange, a diverse set of participants are registered for varied product suites, but the total number of participants was 27.8 million in 2019 (Limaye, 2019) — which is a low number for a large market. Robo-advisory can positively impact this market if it is guided in the right direction by stakeholders and allowed to serve consumers in a fair, equitable way. However, to do that, robo-advisories will have to first address certain key issues related to transparency and accountability of algorithmic decision-making. Building user trust, especially in matters of personal wealth investment, would increase engagement with robo investment advisory services and allow users to reap the benefits they offer.

As mentioned above, a learning system trained on man-made data is likely to pick up some unconscious biases already present in society (Garcia, 2016). Lack of data, biased data, privacy rules, use of wrong tools, irrelevant noisy variables and a number of other reasons cause major problems in these data reliant systems. To laypersons, AI solutions offer very little or no understanding of what happens between the various stages of the process between the input of data and the output of results. Despite the unimaginable computing power and cost reductions, most developers emphasise incrementally improving the performance of AI systems according to a narrowly defined set of parameters and not on how the algorithms are achieving the requisite success. Such challenges

make it difficult to adopt and trust machine learning systems. To assess this risk and regulate algorithms, opening this machine learning 'black box' becomes necessary.

Therefore, two broad approaches have emerged with respect to addressing trust deficit when it comes to artificial intelligence. The first approach attempts to instil human values in AI through a moral code. However, this approach has thrown up complex questions with respect to which value system can be used and how moral and ethical frameworks would translate across cultural boundaries. Further, even if an AI-driven system was instilled with ethical values, the inability to feel emotional consequences in case of failure to abide by those values would continue to render them vulnerable to being bad moral actors. However, solutions like inverse reinforcement learning — where an AI is allowed to observe how people behave in various situations and understand what they value — are said to be showing promise. This approach also brings a broader set of imponderables that are difficult to solve in a quantifiable manner (IBM).

The second approach is to increase transparency by making it easier for individuals to understand decisions being made by AI systems. In fact, industry leaders believe that the technology could get to a point within the next five years where an AI system can better explain why it is recommending certain outcomes to its users (IBM, n.d.). This is at the core of explainability in AI. The idea is to improve users' understanding of how the algorithm is producing results without opening of code or technical disclosures. The Local Interpretable Model-Agnostic Explanations (LIME), for instance, is an algorithm that overcomes AI black boxes. However, with disclosure taking place over an extended period of time, another factor may come into play: behaviour change among users who may game the system by leveraging their understanding of the parameters at play.

The motivation behind this research is to review the current explainability and regulatory landscape, to identify gaps and limitations, and create tools to satisfy both regulatory and user-centric requirements of fintech applications. The research is divided into three broad sections. First, the limitations and practical adoption of algorithmic explainability in an international and Indian context are

evaluated. Second, explainability under the current regulatory standards is explored. An auditing tool is designed to aid regulators without infringing upon intellectual property rights. And third, the benefits of user-centric explainability on user trust and system usability are analysed.

The first research paper in this series, titled "*Trade-off between algorithmic performance and explainability for robo financial advisors in the Indian context*", explores the policy concerns arising out of algorithmic decision making in general, and the fintech and robo-advisory space in particular. The paper is focused on demystifying algorithmic and AI explainability in this particular space and contextualising it to India. A multi-stakeholder interview-based approach would be adopted for this purpose – keeping in mind the views of regulators, firms and users. This would aid in achieving the final aim of providing meaningful explanations about the workings of algorithms, the trade-offs between explainability and algorithmic performance, and the potential harms arising out of incorrect explanations.

The author conducts a review of frameworks developed by a broad spectrum of institutions across multilateral bodies, industry bodies and civil society organisations for assessing algorithms. The ascending trend of explainability as a downstream feature of automated decisions and its potential to ensure accountability and transparency is assessed. This analysis is carried out as an international comparison of algorithmic regulation and governance across five jurisdictions. Finally, the author focuses on the inadequacy of algorithmic regulation in India. This is done by understanding the factors that may distinguish India from other jurisdictions and potential regulatory constraints at play in India. Implementing a right to explainability in India is complicated. At the outset, the reasons for the right to explainability not yet featuring in mainstream legal frameworks on data protection and privacy are many: insufficiency of a standalone right to explainability; lack of consensus on the practical feasibility of providing meaningful explanations; and the potential impact of a legal requirement to implement explainability on regulated businesses.

Interviews are conducted to understand the issue of algorithmic governance from the point of view of the concerned actors from government, regulators and industry professionals. Interviewees delve into the workings of algorithms, the trade-offs between explainability and algorithmic performance, and the potential harms arising from incorrect explanations. Then, the FATE framework is adjusted to the Indian context, considering factors that set India apart from other jurisdictions. These factors include size and population density, linguistic heterogeneity, and income and wealth inequality.

To summarise, the research explores algorithmic explainability in the fintech sector. The authors provide tools to generate explanations that satisfy regulatory and user-centric requirements. The comprehension and perception of these explanations is analysed to determine their effect of user trust in a decision-making system. Finally, the practicality of adoption and limitations of explainability is contextualised to India.

The second paper in the series, *Operationalising algorithmic explainability in the context of risk profiling done by robo financial advisory apps*, is focused on generating explanations for regulatory purposes in compliance with current regulatory practices and requirements. Currently, fintech AI algorithms are not subjected to thorough scrutiny. Due to concerns over confidentiality and intellectual property rights, companies do not provide direct access to algorithms for regulation. In such cases, algorithms can be explained using input and output data. A regulator auditing the algorithm-based on a set of pre-defined regulations or guidelines would increase user trust and ensure that automated investment advisors are unbiased, acting in the user's best interests, and do not face a conflict of interest. With comprehensive and meaningful explanations, regulators could audit the algorithms and check if they comply with the regulations that they are subject to. Algorithms used in automated wealth or investment advisory tools are subjected to Securities and Exchange Board of India (SEBI) regulations in India. However, regulators without technical knowledge possess no means to understand the algorithms and test it themselves. This research aims to develop a 'RegTech' tool with customised explanations that regulators can use to

understand and evaluate the decision-making of any robo-advisory application ADS.

The approach for the 'RegTech' tool is designed using SEBI regulation guidelines for India. In sum, the regulations focus on user risk profiling. Risk profiling in investment advisory algorithms is mandatory, and all investment advice is given based on risk profiling. All fintech tools are required to disclose limitations and suggest mitigations. There are further rules that require them to act in the client's (i.e., the user of the tool) best interests, disclose conflicts of interest, and store data on the investment advice given. The RegTech tool reverse engineers the input-output data to understand how the algorithm takes a decision. Machine learning models are used to recognise and reconstruct three levels of explanations. First, the importance of user inputs on the outcome of the risk profiling algorithm is calculated. Second, relationships between inputs and the assigned risk classes are displayed. Third, decisions for any given user profile are assessed in order to 'spot check' a random data point.

Using this RegTech tool, a comprehensive system audit and inspection of an algorithm is possible, according to the current SEBI guidelines. Further, an explanation for how the algorithm works can be understood through data without direct access to an algorithm. While an explanation for the algorithm is not mandated, the regulator can use this to check if the robo-advisory tool acts in the client's best interest without any unintended machine bias.

The third and final paper in this series, titled "*Algorithmic Explainability in Practice-Evaluating the Effectiveness of Explanations in the Context of Robo Advisory Apps*", focuses on explainability for end users of fintech applications. As discussed above, consumer adoption of fintech applications is hindered due to a lack of trust in their advice and recommendations. Increasing transparency and accountability through regulation is the first step towards AI acceptance. The second step is understanding and satisfying user requirements from decision-making systems.

Most research efforts looking into the explainability of AI takes an "algorithm-centric" view, relying on "researchers' intuition of what constitutes a good

explanation” (Guidotti, 2018). Thus, the result is several varying definitions of an ‘explanation’. Since the research is usually conducted by the machine learning and computer science communities, the focus is on explaining the algorithm's inner workings. Despite emerging solutions to the black box problem, human intervention will be needed to interpret AI decisions. This research aims to improve user trust in AI recommendations by generating user-centric explanations.

The research makes use of the robo advisory use case to design user-centric explanations for a specific high-stakes application to generate and judge high-quality explanations in context. The application gathers user investment preferences and classifies them into a risk category. Based on this category, mutual funds are recommended to users. A review of similar studies under HCI and Cognitive Sciences is used to define user requirements. After which, a review of explainability techniques reveals three approaches to explanation generation (based on scope, availability, and complexity of models). Techniques that satisfy user requirements are used to explain the system. These explanations are then tested using a user study to understand specific and generalisable requirements of ADS users.

The broad objective is to analyse user perception on the usability of the explanations and the whole system. User trust and comprehension are quantified under two different transparencies and complexities of explanations (white vs black box). Moreover, the change in user perception of explanations and system usability is measured in a demographic group membership context. For example, users from different age groups, risk categories, backgrounds, prior robo advisory or investment knowledge, etc. Finally, the study results are analysed for their contributions towards the broader picture of generic guidelines for innovative inclusion.

In sum, this three-paper series, using a range of tools and approaches, examines in detail the under-appreciated regulatory challenges that emerge during the use of algorithms in the field of fintech, and explores ways to resolve them. While the first paper looks into the global standards of AI adoption and the limitations of

explainability in India, the second paper develops tools to generate explanations that satisfy regulatory and user requirements. The final paper then evaluates the effectiveness of user-centric explanations in the context of robo advisory applications. In combination, these three approaches help paint a comprehensive, overarching and nuanced picture of the adoption of AI in fintech.

The results of these three approaches become vital when seen in the context of the rapid rise of artificial intelligence across products, services and industries. If humans are to work alongside machines in this changing world, explainability becomes an important aspect to consider for companies, regulators and users alike, in order for humans to trust the algorithmic systems in place and improve outcomes for all.

Trade-off between algorithmic performance and explainability for robo-financial advisors in the Indian context

Abstract

The rise of automated decision-making systems (ADS) has been remarkable in the past few years, especially so in the field of fintech. This paper is focused on demystifying algorithmic and AI explainability in this particular field, and contextualising the discussion to India. Given the inter-disciplinary nature of this problem, a multi-stakeholder, semi-structured, interview-based approach has been adopted. AI explainability could include explanations on both system functionality (logic behind general operation of the automated system) as well as specific decisions (rationale of particular decisions) within its scope. Through interviews and a detailed review of available literature, this paper examines the current stage of ADS in the Indian financial sector and its potential future, and more importantly, the trade-off between algorithmic performance and explainability, along with exploring the negative effects of incorrect explanations.

Analysis of the interviews suggests that it is not quite a question of *if* ADS will play a significant role in India's financial services sector, but more a question of *when* that will happen, and that regulations will only catch up at a later stage once the penetration has reached a significant level. On the trade-off between explainability and performance, interviewees seemed to be inclined towards explainability being a priority because of their awareness about the quantum of progress that can be made in terms of performance as ADS-driven products become a reality. Overall, the interviews revealed a cautiously optimistic view among stakeholders in terms of increased ADS penetration, its proper regulation and the increase in prominence of explainable AI.

Keywords: Algorithmic decision-making systems (ADS), algorithmic performance, algorithmic explainability, robo financial advisory apps, fintech, financial services, India

Introduction

Section I: Conceptual background

The rapid rise of artificial intelligence and machine learning (AI/ML) across various sectors cannot be denied. Moreover, such automation has meant that human decision making is now being progressively replaced by data-fed algorithms. This trend is visible in a host of domains cutting across the public and private sectors, such as loan approvals for fintech companies, hiring of employees and identification of criminals by law enforcement agencies. These algorithmic tools primarily rely on rich reserves of personal data about individuals, causing a radical disruption in the way decisions have traditionally been made in various arenas.

On a related note, concerns surrounding bias in automated systems have led to a lively debate on the need for explainability of automated decisions. As a means to legally entrench such a notion, the right to explainability has emerged as a potential legal tool to guard against discriminatory outcomes by machines in the real world. However, the debate cannot be considered linear, as it involves a complex web of competing interests and conflicting interpretations of statutory provisions. Therefore, a pursuit of straightforward and obvious solutions is futile, and a better approach would be to appreciate the nuances of each issue.

This research paper focusses on the policy concerns arising out of algorithmic decision-making in general, and the fintech and robo-advisory space in particular. The paper is focused on demystifying algorithmic and AI explainability in field of fintech and robo-advisory, and contextualising it to India. A multi-stakeholder, interview-based approach would be adopted for this purpose, keeping in mind the views of regulators, firms and users.

The 'Introduction' chapter consists of eight sections. Section II, which follows the first and current section on 'conceptual background', will review the debate on algorithmic governance and the rising demand for explainability of automated decisions. Some popular frameworks like FAT/ML, FATES, FATE, ETHICA and AI4SG find mention in this Section, and have been examined in detail in Appendix A to understand their key features and focus areas. Further, the Section will discuss frameworks developed by a wide spectrum of institutions across

multilateral bodies, industry bodies and civil society organisations for assessing algorithms. Most significantly, the chapter delves into the ascending popularity of explainability as a downstream feature of automated decisions, and its potential to ensure accountability and transparency.

Section III undertakes an international comparison of algorithmic regulation and governance across five jurisdictions, namely the United States (US), the United Kingdom (UK), the European Union (EU), Australia and Canada. This study is helpful in analysing the approach adopted by various jurisdictions, and the common conundrum of how best to promote innovation in AI, while at the same time protecting individuals and their personal data. The jurisdictional analysis has been carried out under the buckets of regulatory landscape, algorithmic accountability in particular sectors, regulation of algorithmic bias and the right to explainable AI.

Section IV goes on to study the inadequacy of algorithmic regulation in India. Having traced the extant legal and policy framework consisting of the Information Technology Act, 2000 and the Information Technology (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, this section highlights various sectoral efforts driven by regulators. Moreover, with the Personal Data Protection Bill, 2019 pending before a Joint Parliamentary Committee, it scrutinises the implications of the proposed law for algorithmic decision-making.

Section V looks into the specific needs of the fintech sector, with an eye on robo-advisory, thus enabling the reader to understand the manner in which algorithmic regulation could be appropriately tailored for the sector. The chapter traces extant regulations on fintech automation in India. This includes efforts by the Securities and Exchange Board of India, Reserve Bank of India and the Pension Fund Regulatory Development Authority. The impact of the passage of the Personal Data Protection Bill, 2019 for the fintech sector and its adoption of automated tools has also been covered. The section concludes with recommendations for regulating financial service automation.

In Section VI, the case for implementing a right to explainability in India is studied. At the outset, the reasons for the right to explainability not yet featuring in mainstream legal frameworks on data protection and privacy are analysed. These reasons are as follows: insufficiency of a standalone right to explainability; lack of consensus on the practical feasibility of providing meaningful explanations; and the potential impact of a legal requirement to implement explainability on regulated businesses. Further, the section seeks to situate the right to explainability in an Indian context. This is done by understanding the factors that may distinguish India from other jurisdictions and potential regulatory constraints at play in India.

Section VII sets out to adapt the FATE framework to the Indian context, by highlighting the potential benefits that the widespread adoption of FATE could usher in. Such benefits could be in the form of conceptual cohesion, as well as guidance for algorithm developers and deployers of algorithms for decision-making. The section also discusses possible roadblocks in the implementation of the FATE framework.

Section VIII brings together all concepts necessary to carry out the study, and sets the research undertaken by this paper in context. It looks at the kinds of stakeholders involved, the role of regulators, recognised ways of ensuring responsible AI, ideas around robustness and explainability of AI systems, and the trade-off between explainability and accuracy that is being studied here.

The following chapters go into the research methodology and the analysis of the results. The chapter on research methods puts forth the methodology for conducting interviews with a select group of participants. The interviews are semi-structured in nature, and enable the understanding of the issue of algorithmic governance from the point of view of the concerned actor. A list of potential interviewees from government, regulators and industry have been indicated here. The key questions that would be presented to interviewees involve understanding whether it is possible to provide meaningful explanations about the workings of algorithms, the trade-offs between explainability and

algorithmic performance, and the potential harms arising out of incorrect explanations.

Finally, the chapter on the analysis of the interviews examines the results using five separate themes. It then discusses the concerns surrounding the operationalising of algorithmic explainability in India from an operational perspective. Due consideration is given to factors that may set India apart from other jurisdictions and consequently should inform policy formulation, such as size and population density, linguistic heterogeneity, and income and wealth inequality. After submitting recommendations on setting up of an Indian algorithmic accountability regulator, the chapter offers comments on appropriate regulatory structure, design and approach, followed concluding remarks in the next section.

Section II: Algorithmic governance and the rising demand for explainability

The deployment of artificial intelligence and machine learning in private and public sectors has disrupted the manner in which decisions have traditionally been made. Algorithms have come to mean a variety of things, including having the capacity to shape society, and being “pathways through which capitalist power works” (Ziewitz, 2016). Lee, Resnick and Barton have discussed the rising tendency of private and public sector entities to use AI/ML for automating simple and complex decisions (Lee, Resnick, & Barton, 2019).

Further, the use of such technology has challenged the metrics we use to assess the legitimacy of decisions. In response to these challenges, a wave of ethical, legal and regulatory concerns has emerged on the use of algorithms in decision-making. Thus, algorithms and their deployment can no longer be considered technological issues to be addressed by the engineering community alone. Rather, the intellectual frameworks being formulated should be informed by a wide range of disciplines and stakeholders.

This has prompted the formulation of frameworks to assess the fairness of algorithms. A review of such frameworks is useful in understanding the current

debate on algorithmic accountability, as well as identify the principles gaining precedence in this sphere. Moreover, it is interesting to note that the existence and development of these frameworks have also furthered the demand for explainability of automated decisions. While a few frameworks have overtly emphasised on explainability as a significant component of algorithmic governance, others have done so in an implicit manner. Some popular frameworks, namely FAT/ML, FATES, FATE, ETHICA, and AI4SG, are discussed in detail in Appendix A. They have been analysed on the basis of their origin, core principles and contribution to furthering the debate on algorithmic transparency.

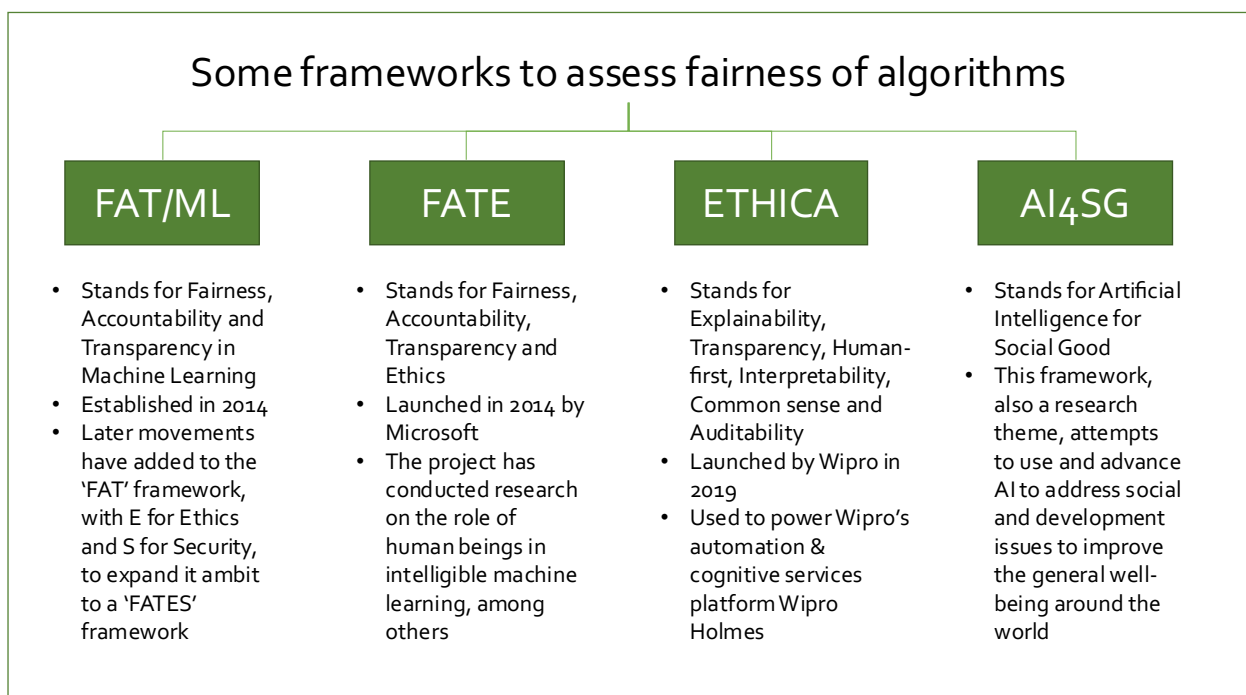


Figure 1. 1 Frameworks discussed in this paper to assess fairness of algorithms

Apart from the select frameworks discussed in Appendix A, there exist other frameworks to assess algorithms developed by multilateral bodies, industry bodies and civil society organisations. It is evident that the movement for responsible and principled use of AI has been steadily gaining traction (Fjeld & Nagy, Principled Artificial Intelligence, 2020). Fjeld et al (2020) have mapped the proposed ethical and rights-based approaches to AI principles across the world (Fjeld, Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI, 2020). Such frameworks have emerged from a wide spectrum of institutions, cutting across the private sector (Google,

IBM, Microsoft), civil society (Amnesty International, Access Now), government (NITI Aayog, European Commission, UK House of Lords), inter-governmental organisations (OECD, Council of Europe, G20) and multi-stakeholder groups (University of Montreal, New York Times).

Their illuminating study has revealed eight key common themes across various AI principles, namely a) privacy, b) accountability, c) safety and security, d) transparency and explainability, e) fairness and non-discrimination, f) human control of technology, g) professional responsibility and h) promotion of human values (Fjeld, *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*, 2020). Interestingly, the authors have also concluded that the principle-based frameworks for AI are beginning to converge, as the more recently mooted frameworks include all eight of the aforementioned themes.

Nevertheless, it is worth questioning the utility of the various frameworks for ethical AI that are now in circulation. The difficulty of using ethical AI frameworks as a benchmark to organise activity, is that the principles contained in these frameworks are often vague, non-binding and unactionable. The lack of a standardised definition leads to ambiguity and places a premium on how a particular principle or framework is interpreted.

In fact, the focus over the recent years on ethics may be construed as a way to sidestep government regulation and instead resort to the non-binding and flexible domain of ethics. (Basu, *What is the problem with 'Ethical AI'? An Indian Perspective*, 2019). According to this view, ethics could be "exploited as a piecemeal red herring solution to the problems posed by AI" (*ibid.*). Thus, it is recommended that the goal to achieve fairness in AI should utilise law as a tool to ensure responsible behaviour, instead of limiting itself to ethics alone. (Basu, *What is the problem with 'Ethical AI'? An Indian Perspective*, 2019) (Mittelstadt, 2019).

It may be expected that there will be a shift to the law as a means to secure responsible AI, through routes such as the legal right to explainability. The

statutory guarantee of such a right would ensure the fulfilment of the key themes discussed above, like accountability, transparency, fairness and non-discrimination, human control of technology, and promotion of human values (Fjeld, *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*, 2020)

The Rising Demand for Algorithmic Explainability

In the four frameworks for assessing algorithms that we mention in the section above (and review later in Appendix A), varying approaches towards explainability may be observed. For example, the ETHICA framework, which is geared towards de-biasing AI models from a functional perspective, has explicitly underlined the need to incorporate explainability into algorithmic governance. Explainability is sought to be deployed as a part of the development process itself, using proper anomaly detection, and human-based auditing. (Wipro, *State of Automation 2019*, 2019).

Further, the FAT/ML framework has considered explainability to be a core principle, alongside responsibility, accuracy, auditability and fairness. The framework also envisages algorithm creators developing a social impact statement based on explainability and the other principles.

While the other frameworks do not explicitly incorporate explainability in their matrix, some such as FATE, implicitly give credence to the need for XAI. In fact, some alternative formulations of the FATE framework have substituted the 'E' for ethics to mean explainability.

The primary reason for seeking explainability is so that individuals are able to comprehend why a particular automated system made a certain decision. The provision of an explanation empowers the individual to assess the adequacy of such explanation – and either agree or disagree with it (Heaven, 2020). For example, explainable AI should help individuals to peek into the black box, and identify which data features are being picked up by a neural network. Consequently, this would aid in understanding whether the resulting model is free

of bias or not (*ibid.*). Explainability becomes important when it is acknowledged that the code is not neutral. In fact, code – or the software and hardware that cyberspace is composed of – determines how easy or hard it is to protect privacy (Lessig, n.d.). Thus, changes in code simultaneously bring about changes in the fabric of cyberspace.

It is evident from the above discussion that the issue of algorithmic regulation throws up thorny questions on fairness, accountability and other values at the heart of a liberal constitutional democracy. In light of this, algorithmic accountability has emerged as a popular area of regulatory demand (Hunt & McKelvey, 2019). Pasquale has urged that public demand for transparency from technology companies that develop algorithms is critical (Pasquale, 2016). This is because it is essential to assess algorithmic decisions for fairness, non-discrimination and openness (Pasquale, 2016).

However, existing legal doctrines may be poorly equipped to tackle algorithmic decision making (Gillis & Spiess, 2019). A central challenge which the automated society must grapple with is as follows: how can our legal frameworks respond to commonplace surveillance and pervasion of algorithmic logic? (Joshi, Welfare Automation in the Shadow of the Indian Constitution, 2020)

The opaque manner in which algorithmic mechanisms function make public assessment difficult. A popular constituent of the demand for algorithmic governance, has been the right to explainability of automated decisions. Therefore, a legally enforceable right for individuals to demand explanations from data-intensive companies would make the latter legally obliged to reveal the reasoning behind complex automated methods.

Wachter, Mittelstadt and Floridi (2017) have discussed the right to explanation of automated decisions, and pointed out that it is viewed as a potentially effective means to demand accountability and transparency in algorithms, AI/ML and automated systems (Wachter, Mittelstadt, & Floridi, 2017). Such explainability would include system functionality (logic behind general operation of the automated system) and specific decisions (rationale of particular decisions)

within its ambit (*ibid.*). However, the authors caution that any right of explanation will encounter significant friction with trade secrets and intellectual property rights (Wachter, Mittelstadt, & Floridi, 2017). Further, the presence of ambiguous legal drafting of the right could render it ineffective (*ibid.*).

Nevertheless, some authors have been sceptical of the efficacy of a right to the logic of automated decisions. Knight has warned that giving users such an explanation may be impossible even for simpler systems like an application showing a targeted advertisement to a particular individual (Knight, 2017). In fact, it may not be possible even for the creators of the applications to fully comprehend or explain the behaviour of algorithms (*ibid.*). It has also been pointed out that the realisation of explainability is domain-dependent (Heaven, 2020). For example, when dealing with complex data like images or text, the neural networks would be relatively deeper and opaquer (*ibid.*). This is worrying because it may require human beings to simply trust in the AI's logic.

It is also a matter of concern that the operation of the right to explainability may hinder the performance of an algorithm. In other words, the easier it is to explain how a machine learning algorithm is working, the less effective the algorithm may be at doing its job. This inverse relationship between the accuracy of a machine learning algorithm and how amenable it is to being interpreted is a significant challenge to efforts at mainstreaming explainability. Several important questions remain unanswered in this sphere. For example, would a shift to explainable AI lead to degradation of system performance? Research on this subject is ongoing, and the goal should be to produce explainable AI that is simultaneously able to deliver a high degree of performance (Turek, n.d.).

On another note, it is worth cautioning that more explanation about automated decisions may not necessarily mean more transparency. A parallel may be drawn to the domain of informational privacy, where tedious and lengthy privacy policies have attained notoriety for their failure to convey digestible information to individuals. In this light, Solove has argued that privacy self-management is unable to provide individuals with meaningful control over their personal data (Solove, 2013). A combination of cognitive problems (uninformed individuals,

skewed decision-making due to bounded rationality) and structural problems (scale, aggregation, assessing harm) make privacy self-management an uphill task. At the individual level, Solove states,

“(1) people do not read privacy policies; (2) if people read them, they do not understand them; (3) if people read and understand them, they often lack enough background knowledge to make an informed choice; and (4) if people read them, understand them, and can make an informed choice, their choice might be skewed by various decision-making difficulties.”

Similar challenges have been observed in other domains like food labelling and consumer protection (Merwe, 2010). As a way to overcome the inherent challenges outlined above, Solove proposes ‘partial privacy self-management’ – a way for individuals to enjoy the empowerment of managing their own privacy, but only till a certain ceiling (Solove, 2013). Beyond this ceiling, it would become an overwhelming task and they would suffer from information fatigue or consent fatigue. Likewise, providing large volumes of information as explanations to automated decisions may not prove to be helpful. Studies indicate that individuals suffer from a particular cognitive bias when it comes to machines: automation bias (Heaven, 2020). The operation of the automation bias instils misplaced confidence in machines despite not understanding the explanations provided (*ibid.*). Thus, providing explanations that can be easily understood by anyone — teachers, police personnel or students — is critical in ensuring a meaningful right to explainability.

Section III: Algorithmic Governance — an international comparison

The apparent and potential harms of algorithmic processing of data have led to calls for regulation. Considering the evolving uses of algorithmic technology, and the lack of a coherent global regulatory framework, various jurisdictions have developed their own approaches.

United States of America

Unlike its prominent counterparts on the other side of the Atlantic, the United States has not enacted a federal privacy or data protection legislation. The

country also does not have a federal data security law. Instead, there exists a mosaic of federal and state legislation intended to protect the personal data of US residents.

The Federal Trade Commission (FTC), an independent federal law enforcement agency, is empowered to protect the interests of American consumers by preventing deceptive, unfair and anti-competitive business practices through enforcement actions. Insofar as privacy overlaps with consumer protection are concerned, it also intervenes to ensure entities are protecting consumers' privacy and personal data. This includes ensuring transparency and meaningful consent, monetary compensation for aggrieved consumers, and deletion of unlawfully procured personal data (FTC, Privacy and Data Security Update: 2018, 2018). In cases of violation of specific privacy laws — for instance, the Children's Online Privacy Protection Act (COPPA), the Fair Credit Reporting Act, and the Controlling the Assault of Non-Solicited Pornography and Marketing (CAN-SPAM) Act — the agency can enforce civil monetary penalties (FTC, Privacy and Data Security Update: 2018, 2018).

Recently, the FTC has issued guidance on AI technology recommending that the deployment of AI-based tools should be "transparent, explainable, fair, and empirically sound, while fostering accountability" (FTC, Using Artificial Intelligence and Algorithms, 2020). It also advocated for explainability, pointing out that if algorithmic processing denies a consumer "something of value", the consumer must be told why (FTC, Using Artificial Intelligence and Algorithms, 2020). In other words, the entity should be aware of what data is used in their model and how it is used to lead to a particular decision.

This guidance was preceded by the 2016 report on big data and the November 2018 hearing on algorithms and predictive analytics (States, 2016). The former advised entities using machine learning and conducting big data analytics to review data sets and algorithms to confirm that hidden biases are not at play in adversely affecting certain populations (FTC, Big Data: A Tool for Inclusion or Exclusion?, 2016). It also cautioned against overestimating mere correlation, and give weight to considerations of fairness and ethics of using big data (*ibid.*). The

2018 hearing acknowledged how little is understood about the evolution of such technologies, and recommended an incremental approach with strong research and development before zeroing in on any policy stance (FTC, Competition and Consumer Protection Implications of Algorithms, Artificial Intelligence, and Predictive Analytics, 2018). The year 2016 also saw the release of the National Artificial Intelligence Research and Development Strategic Plan by the Obama administration to urge scrutiny of algorithms. It served as a guidance and was not legally binding.

Further, there is increasing recognition of the urgency to legally mandate algorithmic accountability and transparency in the US. Thus, the draft Algorithmic Accountability Act, 2019 was proposed by Senators Booker and Wyden, along with Rep. Clarke, to tackle potential bias and discrimination (Act A. A., 2019). This is a significant effort as it constitutes the first federal legislative proposal towards ethical AI. The Bill seeks to mandate entities to fix errant algorithms which are causing unfair, biased or discriminatory decisions, conduct algorithmic impact assessments and data protection impact assessments. The FTC is envisaged as the implementing agency. The Bill is yet to become law.

Apart from federal efforts, states have been spearheading some initiatives on their own. For instance, New York City led the way by enacting an algorithmic accountability law in 2017 (Council, 2017). More recently, legislators in Washington State have also introduced a bill to regulate procurement and use of AI systems in government (Washington, 2019).

Efforts at curbing the spread of facial recognition technology have also gained traction. San Francisco city has imposed an absolute ban on facial recognition technology in order to curb government surveillance through the "Stop Secret Surveillance" ordinance (Francisco, 2019). Similar local proposals to ban facial recognition have emerged in other cities like Somerville and Oakland (Foundation, 2019). Further, the state of Massachusetts has introduced a bill to impose a moratorium on facial recognition and biometric surveillance systems by the state (Massachusetts, 2019).

It can be said that while there are efforts to arrive at a coherent federal regulation on algorithms and AI in the US, it is not clear whether this could be achieved soon. Instead, the framework may remain fragmented for some time to come, with a dispersed mosaic of local instruments in place.

United Kingdom

While there is no dedicated legislation to regulate AI or algorithms in the UK, the country has in place a strong privacy and data protection framework. The Data Protection Act 2018 is an up-to-date version of the erstwhile Data Protection Act 1998. The Act sets out the manner in which personal data may be processed by private and government entities. All entities must follow principles of fair and lawful processing, data minimisation, data retention, purpose limitation and accuracy. Sensitive information such as race, health and biometrics are subject to a heightened degree of protection.

In addition to data subject rights for access, confirmation and portability, the Act also gives individuals rights in case an entity has used their personal data for automated decision-making sans human involvement, or profiling to predict behaviour. The right not to be subject to automated decision-making ensures that a data controller cannot make a decision based on solely automated processing of personal data if it significantly affects a data subject (Act D. P., 2018). A decision would be considered to significantly affect an individual if it has legal consequences (*ibid.*). Further, the law gives individuals the right to intervene in automated decision-making. This right gives individuals the recourse to intervene by requesting reconsideration or relook (Act D. P., 2018). It is available in instances where the data controller has taken a decision that has significantly affected a data subject based solely on automated processing of personal data, and such decision is mandated by law (Act D. P., 2018).

However, the law does not contain a direct, actionable right for individuals to receive information or explanation on automated decisions determined by algorithmic logic (Malgieri, 2019). It can be argued that a wide interpretation of Section 14 of the Act could lead to realising explainability (*ibid.*). The Act and the

EU GDPR discussed below enjoy a complementary relationship, with some provisions of the latter being linked to the former (Act D. P., 2018). Together they constitute the data protection framework for the UK. Other laws in effect such as the Equality Act, 2010 and judicial review under administrative law could also be relevant (Information Commissioner's Office & Institute, Explaining decisions made with AI, 2019).

Apart from the data protection law already in place, the UK Information Commissioner's Office (ICO) has been working towards an AI auditing framework. Last year, the ICO called for inputs on how AI systems can be freed of bias and discrimination (Information Commissioner's Office, Human Bias and Discrimination in AI systems, 2019). It highlighted three technical approaches to mitigate discrimination in ML models, namely anti-classification, outcome and error parity, and equal calibration (Information Commissioner's Office, Human Bias and Discrimination in AI systems, 2019).

The ICO has previously released a guidance in partnership with the Alan Turing Institute on explaining decisions made with AI. The guidance identified six types of explanation (rationale explanation, responsibility explanation, fairness explanation, safety and performance explanation, and impact explanation), and recommended entities using AI to be transparent, accountable, consider context, and reflect on impacts (Information Commissioner's Office & Institute, Explaining decisions made with AI, 2019). Further, it has released a separate guidance on data protection in the context of big data and AI/ML. That guidance recommended entities to adopt privacy by design and embed privacy impact assessment into big data processing to understand necessity and proportionality (Information Commissioner's Office, Big data, artificial intelligence, machine learning and data protection, 2017).

It is interesting to note that in 2017 a House of Lords report on AI in the UK recommended the need for diverse data sets and teams for algorithm development (Lords, 2017). More recently, the Centre for Data Ethics and Innovation (CDEI), an independent advisory board to the UK government, launched an investigation into the potential for bias in algorithms used for

financial services, recruitment and crime prevention (Government, 2019)(Centre for Data Ethics and Innovation, 2019). An interim report of the body highlighted next steps as being gathering evidence and engaging with sectoral stakeholders (Centre for Data Ethics and Innovation, 2019). The final report will be instrumental in determining the shape of financial services using automation in the UK.

The UK does not have laws regulating algorithms or AI/ML at present. However, such regulation could be expected in light of the maturing debate on the harms of algorithmic bias and the need for accountability. Future regulation could be sectoral (financial, employment, law enforcement) as indicated by the CDEI's investigatory approach, or applicable generally. Such regulation would bolster the already existing data protection legislation, and perhaps establish the UK as a leader in AI transparency.

European Union

The EU has already established its position as a pioneer in data protection and privacy. Prior to the EU GDPR coming into force in 2018, the Data Protection Directive of 1995 was in force to regulate the processing of personal data in the EU jurisdiction. At present, the EU GDPR applies to all member states without the need for implementing legislation.

Article 22 of the GDPR provides data subjects the right not to be subject to a decision solely based on automated processing which either has legal effects or significantly affects the data subject. The provision explicitly includes profiling. The provision does not apply in cases where the decision is necessary for a contract, is authorised by law or is based on the data subject's explicit consent. Further, Recital 71 of the GDPR elaborates on the right by stating such decisions may include automatic refusal of an online credit application or online recruiting practices adopted by entities sans any human involvement.

Recital 71 of the GDPR is particularly relevant to algorithmic accountability debates because it mentions the right to explanation ([right] "to obtain an explanation of the decision reached after such assessment and to challenge the

decision"). This is considered to be a suitable safeguard to automated processing along with other safeguards like the right to obtain human intervention, providing specific information to the data subject and not subjecting a child to such processing (GDPR, 2018). Data controllers are encouraged to be fair and transparent, and adopt appropriate technical and organisational measures to minimise risks of error (*ibid.*).

However, unlike the text of Section 22, the recital is not legally binding. Recitals are intended to provide context to the GDPR provisions, and enable faithful interpretation. This means that the right to explanation is not legally binding, as it does not find place in the text of the GDPR itself. This omission has caused much confusion and ambiguity in AI regulation in the EU (Malgieri, 2019)(Wachter, Mittelstadt, & Floridi, 2017)(Edwards & Veale, 2017). To help matters, the Article 29 Working Party has suggested giving a broad interpretation to the scope of Article 22, so that the phrase "solely on automated means" can cover any decision that lacks meaningful human involvement (Party, 2018). Tokenistic human involvement would thus not suffice.

It is worth noting that while member states generally adhere to the text of the GDPR, there exist divergences in the scope of the GDPR provisions and the scope of laws by some member states. This is true in the case of Article 22, with at least four member states (France, Hungary, Austria and Belgium) adopting a wider formulation (Malgieri, 2019).

In 2019, the European Commission's (EC) High-Level Expert Group on AI drew up Ethics Guidelines for Trustworthy Artificial Intelligence. These guidelines set out the requirements of trustworthy AI, namely being lawful, ethical and robust (Commission E. , Ethics guidelines for trustworthy AI, 2019). It also launched a piloting process to engage with stakeholders and test the proposed assessment parameters (*ibid.*).

In 2020, the EC has released a white paper highlighting the European approach to AI, and invited comments for consultation. The 'European approach' indicated therein emphasises on the need to balance competing goals, i.e., promoting AI

innovation in Europe while also supporting ethical and trustworthy AI (Commission E. , White Paper: On artificial intelligence- A European approach to excellence and trust, 2020). For example, the Commission has advocated for a *risk-based approach* to differentiate between high-risk and low-risk AI applications (Commission E. , White Paper: On artificial intelligence- A European approach to excellence and trust, 2020)(Engler, 2020)(Drozdiak, 2020). It has also considered imposing mandatory legal requirements with regard to training data, human oversight and record-keeping(Commission E. , White Paper: On artificial intelligence- A European approach to excellence and trust, 2020).

The EU is an international setter of norms in protecting individual privacy and autonomy. Any regulation stemming from the EU has global implications, with the law having wide reach, and other countries following the EU template. A European AI law could potentially see a repeat of the GDPR, effectively becoming the bible of AI legislation across the world. It is perhaps disappointing to see that despite being a leader, the EU has not given effect to an explicit right to explainable algorithms. This lacuna has been criticised due to its adverse implications for individuals affected by algorithmic processing (Institute, 2018). It remains to be seen whether upcoming AI regulation in Europe seeks to remedy this omission.

Australia

Australia has been an early mover in enacting privacy legislation, with the Privacy Act being in force since 1988. The law has adopted a hybrid co-regulatory model envisaging roles for both industry and government. The Act contains thirteen privacy principles, popularly known as Australian Privacy Principles (APPs). The principles discuss transparent management of personal information, anonymity and pseudonymity, use or disclosure of personal information, security of personal information etc.

However, the law may be considered outdated and needing updating, as it does not adequately account for harms arising from automated processing. The only mention of automation in the Act is in the context of biometric identification

(Privacy Act, 1988). Other legal instruments relevant to the protection of human rights include the Australian Constitution, anti-discrimination laws and international obligations (Commission A. H., Human Rights and Technology: Discussion Paper, 2019).

The Australian government has started responding to the regulatory demands of new technology. For instance, in 2019 the government passed the Consumer Data Right law to bolster customer control over data held about them (Commission A. C., 2017). The law is likely to have a significant impact on the banking sector. Nevertheless, piecemeal efforts of this nature will not be of much use in addressing the challenges of AI.

Recently, the Australian Human Rights Commission has urged the government to modernise the country's privacy and human rights legislation in keeping with the spread of AI/ML (Commission A. H., Human Rights and Technology: Discussion Paper, 2019). The Commission released a discussion paper on the interface between technology and human rights in December 2019. The paper suggested that fundamental tenets of accountability and rule of law should be suitably applied to the evolution of AI (*ibid.*). Extant laws would continue to apply to AI.

However, in cases where AI is causing old issues like unlawful discrimination to emerge in novel avatars, the Commission recommended modernising the Australian regulatory approach towards AI. Moreover, where there exist "problematic gaps in the law", the Commission suggested 'targeted reform' with focus on areas with significant risk of harm, for example, facial recognition technology (Commission A. H., Human Rights and Technology: Discussion Paper, 2019). Implementation of the final report is slated for 2020-21 (Commission A. H., Human Rights and Technology, Our Work).

Previously, the Commonwealth Scientific and Industrial Research Organisation released a study funded by the Australian government, on developing an ethics framework for use of AI in Australia. The discussion paper sought public comments on principles for ethical AI and appropriate tools for the same. Posited

core principles for AI include fairness, contestability, and transparency and explainability (CSIRO, 2019). The proposed toolkit for ethical AI includes tools like impact assessments, internal or external review, and risk assessments (CSIRO, 2019).

The lack of a modern legal and regulatory framework for AI in Australia has led to the adoption of self-regulation among entities. Latest developments indicate that Australia is moving towards legislation to regulate harms arising from automated processing. It would be interesting to observe the approach chosen in this regard, and how it compares with upcoming legislation in the other countries discussed in this chapter.

Canada

The federal privacy and data protection framework in Canada consists of two legislations, namely the Privacy Act, 1983 and the Personal Information Protection and Electronic Documents Act (PIPEDA), 2000. The former is relevant for interactions between the individual and the federal government. The Privacy Act protects personal data in possession of government bodies. Under the Act, individuals enjoy the right to access their personal data as held by the government. For example, it applies in case of provision of public services like old age benefits, employment insurance and tax refunds.

PIPEDA concerns private sector entities in Canada that collect and process personal data during a commercial activity. Commercial activities include any transaction, act or conduct of a commercial nature (Canada O. o., PIPEDA in brief, 2019). Under the law, individuals enjoy the right to access their personal data, challenge its accuracy, and expect their data to be kept secure. Private sector organisations are expected to comply with the 'PIPEDA fair information principles' that include accountability, accuracy and completeness, openness and limiting use, disclosure and retention (Canada O. o., PIPEDA fair informational principles, 2019).

Similar to Australia, the Canadian law also follows a co-regulatory approach with cooperation between industry and government (Srikrishna, White Paper of the

Committee of Experts on a Data Protection Framework for India, 2017). Apart from federal privacy legislation, some provinces like Alberta, Quebec and British Columbia have enacted their own privacy laws applicable to the private sector (Canada O. o., PIPEDA in brief, 2019).

In 2017, the Information and Privacy Commissioner of Ontario recognised that government use of big data has thrown up contested ethical questions, and thus released a set of guidelines for the same. The guidelines seek to guide entities on the best practices to follow while using big data. Dividing the four stages of conducting big data projects into collection, integration, analysis and profiling, the guidelines discussed the operation of ethical AI at each level (Ontario, 2017). It posited some best practices, such as publishing a description of their big data project on the institution website, giving the same treatment to publicly available personal data and non-publicly available personal data, and ensuring that analysed information is accurate, complete and up-to-date (Ontario, 2017). However, it did not envisage realising a right to explainability.

In 2019, the Canadian government passed a Directive on Automated Decision-Making to make sure that deployment of AI is done in consonance with fundamental principles of administrative law like transparency, legality and procedural fairness. The Directive has been issued under powers contained in the Financial Administration Act, and the Policy on Management of Information Technology. It has defined 'automated decision systems' as including "any technology that either assists or replaces the judgement of human decision-makers" (Canada G. o., Directive on Automated Decision-Making, Appendix A, 2019).

The Directive's expected results include data-driven decisions of the federal government that are also procedurally fair, assessment of algorithmic impact on administrative decisions and data on use of automated systems in government institutions to be made public (Canada G. o., Directive on Automated Decision-Making, Appendix A, 2019). To this end, the Directive has designated the Assistant Deputy Minister to conduct algorithmic impact assessments before any automated decision system is produced (*ibid.*). The algorithmic impact

assessment contains a list of sixty questions related to business practices, technical systems and decision oversight (Canada G. o., Algorithmic Impact Assessment, 2020).

Further, the Canadian government released a Digital Charter, and slew of proposals to update the PIPEDA according to modern needs. The ten principles contained in the Charter (for instance, universal access, control and consent, and data for good) are intended to establish Canada as a leader in the digital economy, and modernise norms pertaining to the digital domain (Canada G. o., Minister Bains announces Canada’s Digital Charter, 2019). This year, the Office of the Privacy Commissioner of Canada has called for views on proposals for appropriate regulation of AI. The proposals are about the appropriate definition of AI within the law, adopting a rights-based approach, and creating rights against automated processing (Canada O. o., Consultation on the OPC’s Proposals for ensuring appropriate regulation of artificial intelligence, 2020).

These policy trends indicate that Canada is well aware of the need to bring its existing laws up to date to the challenges posed by AI, and usher in new regulation with tools such algorithmic impact assessment. The manner in which the PIPEDA will be amended remains to be seen, along with whether we will see a federal law on algorithmic accountability. Further, whether Canada will give its residents an explicit right to explainable AI is an open question.

Table 1. 1 Analysis and comparison of the approach adopted by various jurisdictions

Jurisdiction	Federal privacy law	Relevant algorithmic accountability regulation	Scope of algorithmic accountability regulation	Sectors or uses in focus	Status of explicit right to explainability
USA	N/A	“Stop Secret Surveillance” ordinance, San Francisco city Local proposals in	Use-case specific; territorial limitation to city or state areas.	Consumer protection and welfare; financial services; deployment of AI-based tools by government	Absent

		Somerville, Oakland to ban facial recognition		agencies; misuse of facial recognition technology and biometrics	
		State of Massachusetts bill to impose moratorium on facial recognition and biometric surveillance			
UK	Data Protection Act, 2018	Data Protection Act, 2018	Sector-blind	Financial services, recruitment and crime prevention	Limited
EU	EU GDPR, 2018	EU GDPR, 2018	Sector-blind	Sector-blind and risk-based approach	Limited
	Member state laws implementing GDPR	Member state laws implementing GDPR			
Australia	Privacy Act, 1988	Consumer Data Right law	Sector-blind	Consumer autonomy	Absent
Canada	Privacy Act, 1983	Directive on Automated Decision-Making, 2019	Sector-blind	Sector-blind approach	Absent
	Personal Information Protection and Electronic	Digital Charter, 2019			

The above table provides a snapshot of the comparative study assessing five jurisdictions, namely the USA, UK, EU, Australia and Canada. The parameters for comparison have been chosen with the following goals in mind: a) to understand the maturity of existing legal and regulatory frameworks on privacy and data protection, and b) to understand the maturity of algorithmic accountability regulation. In light of the two goals outlined, the parameters chosen were: a) federal privacy law, b) relevant algorithmic accountability regulation, c) scope of algorithmic accountability regulation, d) sectors or uses in focus, and e) status of explicit right to explainability.

The comparative review of frameworks in these five jurisdictions indicate that while dealing with AI accountability, countries are faced with a uniform struggle – how to sufficiently equip old laws to respond to new and emerging threats arising from automated decision-making. They are also faced with the conundrum of how to promote innovation in AI, while at the same time protecting individuals and their personal data. Further, it is plausible that the fear of trampling business innovation has deterred governments from taking a strict stance on algorithmic accountability in haste. It is worth noting that a universal regulatory benchmark has not yet been arrived at.

It will be interesting to observe which approach is ultimately favoured by each jurisdiction, and the brass tacks of how they secure individual liberty. For instance, the EU appears to be driving at a risk-based approach, while Canada seems to favour a rights-based approach. Further, potential divergences in regulation could have implications for an increasingly inter-connected world economy. On the whole, it may be said with some degree of certainty that the next few years will see a surge in law and policy making to regulate harms arising from AI/ML, and automated decision-making in particular.

Regulatory landscape

A significant determinant of the regulatory approach in each jurisdiction is whether it has extant federal privacy legislation. In this regard, the USA is an outlier. While the UK (The Data Protection Act, 2018, EU GDPR, 2018), the EU (EU GDPR, 2018), Australia (Privacy Act, 1988) and Canada (Privacy Act, 1983) have federal data privacy laws in place, the USA does not. This is not to say that the presence of a data privacy law has meant up-to-date algorithmic accountability regulation. Nevertheless, such a law has a stabilising effect and serves as the lodestar in grounding further regulatory developments in emerging areas of technology policy like algorithmic accountability.

For example, the UK, EU and Canada have used their dated federal privacy laws to serve as a base in passing more up to date legislation that is better suited to present times. This may be observed in the UK's move to phase out its older Data Protection Act, 1998 and usher in an updated version that is consistent with the EU GDPR. It is interesting to note that the updated legislation has wide similarities to its predecessor, which could be said to have shaped the UK's culture on privacy and data protection since 1998.

Similarly in the EU, the older version of the GDPR (the Data Protection Directive) had been in force since 1995. It thus served as the foundation for the present GDPR 2018 that is attuned to modern needs in data protection. Canada has followed a similar trajectory, with the Personal Information Protection and Electronic Documents Act, 2000 following the older Privacy Act, 1983. However, the former has not replaced the latter, and both continue to operate simultaneously.

The effects of absence of a federal law on data protection may be observed in both the USA and India, as well as Australia to a limited extent. In the US, the gap has been filled up by a mosaic of federal and state laws seeking to protect various aspects of personal data of US residents, with the FTC being a significant authority. However, there is a noticeable lack of coherence in such efforts as each of them pertain to narrow issues such as misleading advertisements or credit reporting. In the case of India, the absence of a central data protection law has led

to a fragmented regulatory framework with various instruments having sector-specific applications.

Australia has had a slightly different experience, where its dated privacy law has proven to be inadequate. The Australian Privacy Act, 1988 is still in force but the government is likely to pass an updated version in the future. Interestingly, the lack of a modern data protection law has encouraged the adoption of self-regulatory strategies among entities.

On a review of the above jurisdictions, some inferences may be made. First, the absence of a federal privacy law may hinder the development of algorithmic accountability regulation. This is because, as seen in the case of the UK, EU and Canada, a federal privacy law plays a foundational role on which updated regulations on algorithmic accountability could be built. Without that substratum, even specific laws on algorithmic accountability may face difficulties in operation. The draft Algorithmic Accountability Act, 2019 proposed in the USA could expect such obstacles in its course.

An additional difficulty in such a scenario would be the potential conflicts in standards imposed by various laws. For example, there may be conflict between federal laws on algorithmic accountability, as well as between federal and state laws on algorithmic accountability. Such conflicts may lead to regulatory uncertainty for business, as well as violation of individual rights.

Algorithmic accountability in particular sectors

Given the nascence of dedicated regulatory frameworks for algorithmic accountability having a sector-blind ambit, it is useful to examine which sectors have seen the most movement on algorithmic accountability regulation. In this regard, the UK Centre for Data Ethics and Innovation's investigation into algorithmic bias is particularly significant. The CDEI is an independent advisory board to the UK government, and chose to devote attention to algorithmic bias in financial services, recruitment and crime prevention. These sectors were chosen due to their capacity to make decisions that significantly affect the lives of individuals (Centre for Data Ethics and Innovation, 2019).

In the USA, most FTC regulations on privacy and data protection have pertained to consumer protection and welfare. However, these regulations do not appear to regulate algorithmic accountability. The FTC has taken note of the need to regulate algorithms, and recently issued a guidance on AI technology. The guidance has recommended that the deployment of AI based tools should be “transparent, explainable, fair, and empirically sound, while fostering accountability” (FTC, Using Artificial Intelligence and Algorithms, 2020). The US House Financial Services Committee has opted to study measures to reduce bias in automated financial services (Lofchie, 2020).

Moreover, there has been widespread concern about the deployment of AI-based tools by US government agencies. Hence, a bill introduced in Washington State sought to regulate procurement and use of AI systems in government (Washington, 2019), and San Francisco city has gone so far as to impose a complete ban on the use of facial recognition technology with a view to curbing government surveillance (“Stop Secret Surveillance” ordinance)(Francisco, 2019). Local efforts along the same line have sprung up in Somerville, Oakland the state of Massachusetts (City of Somerville, 2019)(Foundation, 2019)(Massachusetts, 2019).

The EU is following an ordered process of coming up with a regulatory framework for algorithmic accountability. Recently, the European Commission’s High-Level Expert Group on AI floated the Ethics Guidelines for Trustworthy Artificial Intelligence. The guidelines have proposed certain pre-requisites of trustworthy AI, namely being lawful, ethical and robust (Commission E. , Ethics guidelines for trustworthy AI, 2019). It is likely that as opposed to a fragmented approach of regulating specific sectors, the EU would opt to first erect a sector-blind regulation for algorithmic accountability. The EC’s white paper on AI indicates that such a framework would be informed by a risk-based approach to differentiate between high-risk and low-risk applications of AI (Commission E. , White Paper: On artificial intelligence- A European approach to excellence and trust, 2020).

Australia has chosen to view algorithmic accountability through the lens of consumer autonomy, and passed the Consumer Data Right law to bolster customer control (Commission A. C., 2017). It is likely to have an impact on the banking sector in particular. Further, Canada appears to be driving at strengthening its larger regulatory framework for use of AI in order to establish itself as a leader in the digital economy. Therefore, it has focussed its attention on sector-blind regulations such as the Directive on Automated Decision-Making, 2019 and the Digital Charter, 2019.

The Regulation of Algorithmic Bias and Right to Explainable AI

Algorithmic accountability may be achieved by empowering individuals with actionable rights. Such rights enable individuals to assert control over their personal data, demand accountability and seek redressal from entities deploying AI. While the right to explainable AI has emerged as a popular demand in this debate, it is worth noting that a meaningful formulation of the right is presently lacking in legal frameworks across the world.

The traditional rights framework for individuals has been drawn up by the EU GDPR. According to the GDPR, a data subject has the rights of access, rectification, to be forgotten, restriction of processing, data portability and objection. Additionally, the EU GDPR guarantees data subjects the right to not to be subject to a decision based solely on automated processing under Article 22. However, this provision does not mention the right to explainability in its text. Instead, mention of the right is found in Recital 71 of the GDPR.

Recital 71 posits the right to obtain an explanation of an automated decision as a safeguard to automated processing. This is envisaged as one among other safeguards including the right to obtain human intervention, providing specific information to the data subject and not subjecting a child to such processing (GDPR, 2018). However, the difficulty in legally giving effect to the right to explainability arises from the non-binding nature of the recitals in the GDPR universe. This has led to much confusion among scholars and practitioners, along with criticism on the inadequate scope of the right to explanation contained in

the law. Nevertheless, the global move towards demanding a concrete right to explainable AI is undeniable.

Since the UK follows a similar framework as the EU GDPR, the position is similar to Europe. The UK Data Protection Act, 2018 does not contain an explicit right for individuals to receive information or explanation on automated decisions determined by algorithmic logic (Malgieri, 2019). Thus, akin to the EU GDPR, the matter of whether a right to explanation exists, and the scope of such a right, has become a subject of interpretation. While a broad interpretation of section 14 of the Act could mean the existence of a right to explanation, a narrower interpretation may negate it.

In the USA, the FTC has supported the right to explainability in the interests of consumer welfare (FTC, Using Artificial Intelligence and Algorithms, 2020). The proposed Algorithmic Accountability Act does not appear to impose any binding legal obligation on entities to provide explanations to individuals of decisions arrived at through automated processing. In Australia, enthusiasm for a right to explainability has come from the Data61 discussion paper titled Artificial Intelligence: Australia's Ethics Framework, as well as the Commonwealth Scientific and Industrial Research Organisation study on ethical AI funded by the Australian government (CSIRO, 2019).

Canada makes for a slightly more nebulous case, as it is not entirely clear whether the government is keen on carving out a legal right to explainability. In fact, the Information and Privacy Commissioner of Ontario in 2017 released guidelines for ethical questions in the use of big data. While the guidelines contained discussion of best practices to follow while using big data, it did not envisage a right to explainability for individuals. This is surprising considering Canada's efforts to become a global leader in the digital economy.

Nevertheless, Canada has in place a robust system to conduct algorithmic impact assessments under the Directive on Automated Decision-Making, 2019. Further, the call for views by the Office of the Privacy Commissioner of Canada on proposals for appropriate regulation of AI contains a discussion on the adoption

of a rights-based approach and the creation of rights against automated processing (Canada O. o., Consultation on the OPC's Proposals for ensuring appropriate regulation of artificial intelligence, 2020). Since the right to explainability would form a crucial part of a rights-based framework against automated processing, it is hoped that Canada will eventually accord it sufficient importance.

Section IV: Review of algorithmic regulation in India

Extant legal and policy framework

In sync with the struggle of the jurisdictions discussed above, India too has been grappling with the question of how best to modernise its legal framework to effectively respond to automation. The Indian constitutional and administrative law paradigm is furnished towards assessing the rationality and legitimacy of decisions made by human actors. Thus, it is a struggle to make them apply neatly to decisions made by non-human actors.

We have observed above that the UK, EU, Australia and Canada share a common characteristic, i.e., they have an extant privacy or data protection legislative framework. In contrast, India does not have a privacy or data protection in place at present. The extant legal framework is limited in scope, with the primary legal instrument being the Information Technology Act, 2000 (IT Act) and the Information Technology (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011 (SPDI rules).

The SPDI rules were issued under Section 43A of the IT Act, and set out reasonable security practices and procedures to be implemented by body corporates. However, the SPDI rules are limited in their scope as they only regulate sensitive personal data (like passwords, financial information, and medical records and history) and not personal data in general. Further, they only cover private entities and not to the State. Apart from the above two instruments, norms having implications for data privacy are contained in sectoral statutes pertaining to taxation, right to information and banking. Thus, at present there is

a lack of uniformity and direction in the regulatory framework pertaining to privacy.

Neither the IT Act nor the SPDI rules were drafted to respond to the challenges posed by AI/ML or algorithmic processing of personal data. Hence, it is not surprising that such existing laws have proved to be poorly equipped in addressing novel challenges like bias and discrimination arising from automated processing of personal data, and have thus failed to meet the consequent need for explainability.

Apart from the general legal framework discussed above, it would be useful to examine some sectoral practices in the financial sector. In this regard, it is worth noting that it is crucial to study the financial sector, due to its unique position in affecting lives of ordinary individuals. The financial sector has traditionally used data to support decision making and arrive at accurate predictions about behaviour of individuals (Centre for Data Ethics and Innovation, 2019). For instance, financial services organisations may need to assess the likelihood of a certain individual repaying their debts in a timely manner. However, it is essential to look at the financial sector through a critical lens, due to its legacy issues of historically underrepresenting particular groups, and to ensure that historic biases are not further perpetuated (Centre for Data Ethics and Innovation, 2019).

In the absence of an adequate privacy law, some sectoral regulators were prompted to set their own standards. In a 2002 report, the Reserve Bank of India (RBI) indicated the significance of maintaining information security, with the preservation of confidentiality, integrity and availability of information (RBI, Annexure: Information Systems Security Guidelines for the Banking and Financial Sector (Part 1 of 2), 2002). It recommended how information can be secured in the banking and financial system.

In a 2017 report on household finance, the RBI preferred a rights-based approach to privacy as opposed to the standard consent-based model. In doing so, the regulator noted how AI/ML and big data have changed the nature of data processing, and thus weakened the efficacy of consent as a tool to guard privacy

(RBI, Report of the Household Finance Committee, 2017). The Committee also pointed out that “algorithmic provision of household finance services” could lead to discrimination “if left unchecked” (RBI, Report of the Household Finance Committee, 2017).

Recently, the RBI released a circular requiring all data relating to payment systems to be stored locally in systems within the territory of India within six months. The circular would be applicable to end-to-end transaction details, information collected and processed, and payment instruction (RBI, Storage of Payment System Data, 2018). This strict localisation mandate stemmed from the need to institute safety and security measures for the vast volumes of payment data being processed (RBI, Storage of Payment System Data, 2018). In sync with the larger pro-data localisation trend in the country, the RBI believed that local storage of data would enable smoother monitoring and investigation. The present move towards account aggregators (non-banking financial companies that manage consent for financial data sharing), with the RBI giving out licenses to few entities, also indicate the RBI’s urge to strengthen data protection and consent.

Similarly, the Securities and Exchange Board of India (SEBI) has previously circulated guidelines for seeking data to boost data analytics, research and academic studies. It was intended to lend coherence to the process of data sharing and formalise data protection measures, in order to prevent misuse and unauthorised access. It required data seekers (like educational institutions, research organisations and other regulators) to sign an undertaking of confidentiality and non-disclosure (SEBI, Guidelines for Seeking Data, 2019). The guidelines also specified that only data that is at least two years old would be eligible for sharing.

The Insurance Regulatory and Development Authority of India (IRDAI) has been particularly active on this front. With rapid digitisation of the insurance sector, the IRDAI has stepped up by specifying an additional framework to safeguard personal data of policyholders. Insurance companies have been mandated to maintain confidentiality of collected policyholder information (IRDAI, (Protection

of Policyholders' Interests) Regulations, Regulation 19(5), 2017). Insurers must also ensure that storage systems have adequate security features, records are stored in local data centres in India, and data shared with outsourced service providers remains confidential and secure. The IRDAI has also formulated cybersecurity guidelines to ensure that entities implement measures for confidentiality, integrity, and consistency of data in a systematic manner (IRDAI, Cyber Security Guidelines, 2017). A separate set of e-commerce guidelines were also issued to bolster e-commerce in the insurance space and lower cost of transactions (IRDAI, Guidelines on insurance e-commerce, 2017).

Personal Data Protection Bill, 2019

The most critical judicial development on privacy in India has been the decision of the Supreme Court in *Justice K.S. Puttaswamy v. Union of India* ('Puttaswamy'). In this case, the Court was faced with the question of whether a fundamental right to privacy exists or not. The judgment accordingly established a fundamental right to privacy under the Constitution of India. The Court unambiguously held that privacy flows from the right to life and personal liberty under Article 21, as well as other fundamental rights contained in Part III of the Constitution to the extent that they intersect with autonomy and dignity (*K.S. Puttaswamy v. Union of India* (majority opinion delivered by Chandrachud J.), 2017).

The right to privacy was considered to lie across a spectrum of existing rights, and thus giving it explicit recognition does not amount to a constitutional amendment (*K.S. Puttaswamy v. Union of India* (majority opinion delivered by Chandrachud J.), 2017). Three concomitant facets of privacy, namely bodily privacy, informational privacy, and privacy of choice were recognised.

As informational privacy is a facet of the larger right to privacy, the bench urged the Union government to put in place a framework for personal data protection. Such a regime would strike an appropriate balance between individual interests and legitimate concerns of the State. Legitimate concerns of the State include national security, prevention and investigation of crime, allocation of resources for human development, revenue, encouraging innovation, and preventing the dissipation of social welfare benefits (*K.S. Puttaswamy v. Union of India* (majority

opinion delivered by Chandrachud J.), 2017). The need for a data protection law was emphasised again by the Supreme Court in its judgment on the constitutionality of the Aadhaar scheme (K.S. Puttaswamy v Union of India, 2019).

Further, a three-part test was established by the Court to assess the constitutionality of any privacy invasion. First, the privacy restricting measure should be backed by a law. Second, as a safeguard to arbitrariness, the restraint must be in pursuance of a legitimate state aim that qualifies as reasonable under Article 14 of the Constitution. Finally, the means chosen by the restraint should be proportionate to avowed goal of the law (K.S. Puttaswamy v. Union of India (majority opinion delivered by Chandrachud J.), 2017).

Close on the heels of the judgment, the Committee of Experts under the Chairmanship of Retd. Justice B.N. Srikrishna released the draft Personal Data Protection Bill, 2018 and report on a free and fair digital economy. Subsequently, the Bill was introduced in the winter 2019 session of the Lok Sabha with some changes made. Currently, a Joint Parliamentary Committee (JPC) under the chairpersonship of Ms. Meenakshi Lekhi is reviewing the Bill, and sought views and suggestions from stakeholders. It is expected that the Bill will be passed soon by the Parliament.

The passage of the Bill will effectuate a radical shift in the privacy and data privacy ecosystem in India. At the outset, the proposed law will substitute section 43A of the IT Act and the SPDI rules. Moreover, it will create a framework for imposing obligations on data fiduciaries and securing the rights of data principals. It is interesting to note that the Bill has chosen to view the relationship of personal data collection and processing through the lens of trust. Here, the entity who determines the purpose and means of processing personal data is considered the 'data fiduciary' (Personal Data Protection Bill, 2019). The data fiduciary could be any person, including the State, company, juristic entity or individual. 'Data principal' refers to the individual to whom the personal data relates to.

The legislative intent of the fiduciary relationship reflected by the Bill was explained in the Srikrishna Committee's report on a free and fair digital economy. According to the report, the relationship between the data processing entity and the individual is built on trust (Srikrishna, A Free and Fair Digital Economy: Protecting Privacy, Empowering Indians, 2018). Thus, an individual expects her personal data to be processed in a fair manner that is respectful of her interests. This places an onus on entities to honour a duty of care in fair and responsible processing of data for reasonably foreseeable purposes (Srikrishna, A Free and Fair Digital Economy: Protecting Privacy, Empowering Indians, 2018).

The Bill has several implications for algorithmic processing of data and potential harms flowing therein. Since the law will apply across the board, it will bind both the state and private entities to fiduciary obligations. Data fiduciaries will henceforth be bound by principles of purpose limitation, fair and reasonable processing, collection limitation, notice and consent, data quality, data storage, and accountability (Personal Data Protection Bill, 2019). Sensitive personal data, such as passwords, financial data and biometric data are subject to heightened protection.

Data fiduciaries are prohibited in processing personal data lacking a specific, clear and lawful purpose (Personal Data Protection Bill, 2019). Personal data must be processed in a fair and reasonable manner to ensure the privacy of the data principal, and for such purpose as consented to be the data principal. Incidental or connected purposes which the data principal would have reasonable expected her personal data to be used for are also permitted (*ibid.*). Processing should account for the context and circumstances in which collection of personal data took place (Personal Data Protection Bill, 2019).

Any algorithmic processing would have to comply with the above obligations imposed. This may prove to be challenging for data fiduciaries and processors, especially for big data. The Srikrishna Committee noted in its report that big data processing has emerged as a "frontal challenge to the well-established principles of collection limitation and purpose limitation" (Srikrishna, A Free and Fair Digital Economy: Protecting Privacy, Empowering Indians, 2018). Principles of collection

limitation and purpose limitation seek to circumscribe the boundaries of a particular data processing activity.

In contrast, big data processing is predicated on collection of vast volumes of personal data at scale and subsequently identifying appropriate uses. As some uses only become clear after the combination of personal data collected from different sources, it would be difficult to communicate them to the data principal at the time of collection. Thus, the report aptly observed that “meaningful purpose specification is impossible with the purposes themselves constantly evolving” (Srikrishna, *A Free and Fair Digital Economy: Protecting Privacy, Empowering Indians*, 2018). To remedy this, the Bill has narrowly tailored use of big data analytics by restraining its uses through various provisions. For instance, while the Bill contains an exemption for personal data processed for research purposes, anonymisation is required as far as possible. Moreover, a general duty has been imposed on researchers to make sure that individuals are not harmed or targeted by the research.

Moreover, by enabling individual data principals to establish control over their personal data, the Bill puts in place measures that will curtail the adverse effects of automated data processing. Individuals will enjoy the right to confirmation and access, correction and erasure, data portability, and to be forgotten (Personal Data Protection Bill, 2019). This means that individuals will be empowered to obtain information from the data fiduciary and have remedies for unlawful processing.

The rights guaranteed to data principles also have implications for data processing for profiling (Joshi, *India’s privacy law needs to incorporate rights against the machine*, 2020). Profiling has been defined under the Bill as any processing of personal data involving analysis or prediction of behaviour, attributes or interests of the data principal. This indicates that the Bill has been drafted based on an understanding of automated processing and profiling of individuals.

With regard to the right to explainability, the Srikrishna Committee in its White Paper submitted that it may not be appropriate to merely mandate providing the logic for automated decision making by law (Srikrishna, White Paper of the Committee of Experts on a Data Protection Framework for India, 2017)(Srikrishna, White Paper of the Committee of Experts on a Data Protection Framework for India, 2017). Instead, it advocated for a harm-based approach, suggesting that individuals should be protected against the harms arising from such decision making. This could only be realised through rights that are both legally tenable as well as feasible contained in a broader data protection law (Srikrishna, White Paper of the Committee of Experts on a Data Protection Framework for India, 2017).

Thus, the Committee favoured putting in place provisions for internal and external audits in organisations that deploy algorithmic decision making for significant amounts of personal data. This would ensure accountability of the data fiduciary, and necessitate the maintenance of robust records to comply with the data protection law (Srikrishna, White Paper of the Committee of Experts on a Data Protection Framework for India, 2017).

On the right to object to automated processing and access the logic behind it, the Committee decided against including it in the data protection law. It expressed approval for an ex-ante accountability framework for certain data fiduciaries which engage in evaluative automated decisions (Srikrishna, A Free and Fair Digital Economy: Protecting Privacy, Empowering Indians, 2018). This can be part of proactive compliance for privacy by design by data fiduciaries, which could be regularly audited and monitored by the data protection regulator.

While data principals have been guaranteed rights to have a say in such automated processing, the Bill has some noticeable gaps. Individuals have not been given explicit protection against particular harms arising from automated decision making or profiling. For instance, the right to explainability of automated decisions has not found a place in the Bill. Giving legal recognition to an individual's right to explainability would have made existing obligations placed on data fiduciaries (such as fair and reasonable processing) more meaningful. The

finer details of carving out additional safeguards or restrictions against profiling has been left to the regulation-making power of the Data Protection Authority (DPA). Even this will apply only to a sub-category of personal data, i.e., sensitive personal data, and not all personal data (Personal Data Protection Bill, 2019).

Furthermore, the Bill provides especially weak protection to harms arising from automated decision making done by the State. Broad exceptions have been given to the State for processing of personal data. For instance, the State does not require consent for processing personal data necessary for “any function of the State authorised by law” for “the provision of any service or benefit to the data principal from the State” or “the issuance of any certification, licence or permit for any action or activity of the data principal by the State” (Personal Data Protection Bill, 2019). The requirement of necessity is heightened to “strict necessity” for processing of sensitive personal data by the State.

The set of provisions setting out state exceptions from consent curtails individual control over processing of personal data by the State (Marda, 2018). In light of pervasive adoption of AI/ML and automated decision-making systems by the State, the provisions of the draft law appear inadequate. The standard of strict necessity specified for sensitive personal data may not be useful, as machine learning systems are trained on vast datasets consisting of both. Such systems may absorb and exaggerate bias and discrimination from these datasets (Marda, 2018). The Bill does not provide effective remedy against such potential harms.

While the Bill has made strides in securing individual autonomy over their personal data, and created a framework to hold data processing entities accountable, it has held back on addressing challenges posed by algorithmic decision-making. It is hoped that the JPC reviewed version of the draft law will fill these lacunae and ensure better realisation of individual autonomy against automated decision-making.

Section V: Contextualising algorithmic regulation to fintech & robo financial advisors

While Section III conducted a comparative study of the jurisdictions in US, UK, EU, Australia and Canada to provide a global snapshot of contemporary algorithmic practices, Section IV reviewed the relevant legal regulations in place around the overall AI landscape in India. Now, Section V attempts to focus this discussion specifically around the financial services sector. There is value in studying aspects of the right to explainability in fintech *in India* because of the unique operational characteristics prevalent in the country. For instance, India has the world's second highest national population and is getting increasingly population-dense, requiring setting up of increasingly decentralised regulatory mechanisms. India is also linguistically diverse, and has strong socio-economic cleavages that maintain wealth and social inequalities in its societies. Such factors, which Section VI evaluates in detail, make situating this discussion around algorithmic explainability in India important. This Section, for now, first looks at automation in the fintech sector globally, before exploring the regulation of financial automation in India specifically.

Automation in the financial services sector

The financial services sector provides an appropriate example of the pervasion of AI/ML in decision-making, and the potential harms associated with such automation. For instance, a PwC Global Fintech Report has noted that efforts by UK financial services firms to implement robotic process automation is setting global trends (PwC, 2019). The adoption of automated processing in the sector has meant faster and more efficient decisions, while bringing down the costs of financial services and products.

Algorithmic processing can now be seen in fraud detection, trading decisions, evaluation of loan applications, determination of insurance premiums, etc. Additionally, the future uses of algorithms in the financial services sector are promising. It may thus be concluded that the financial services sector is moving towards robo-advisory strategies, or at least a hybrid of human and robo-advisory strategies (PwC, 2019).

Robo-advisory applications are web-based investment advisory algorithms that are automated, and designed to recommend "the best plans for trading,

investment, portfolio rebalancing, or tax saving, for each individual as per their requirements and preferences” (Krishnan, Deo, & Sontakke, 2020). The predictions are generally based on the client filling up a questionnaire and subsequently being categorised on a spectrum of risk ranging from low to high (*ibid.*). This is then used to dispense financial advice for a range of goals such as retirement, education or a rainy-day fund (Bank, 2019). Robo-advisers are considered to be cheaper and more accessible *vis-a-vis* human advisors, and could democratise financial services by providing financial advice to hitherto unbanked segments of the population (Krishnan, Deo, & Sontakke, 2020).

This has also meant that the potential harms associated with algorithmic processing, such as bias and discrimination, are likely to play out in the financial services sector too. Financial firms enjoy significant sway over the lives of individuals through their decisions. As discussed above, the UK’s CDEI (an independent advisory body formed by the UK government) took cognisance of such pervasion and announced a review of algorithmic bias in four key sectors, including the financial services sector. The House Financial Services Committee in the US has also undertaken a similar exercise to identify measures to reduce bias in automated financial services (Lofchie, 2020).

Explaining the rationale behind the selection of these particular sectors, their interim report stated that these sectors are engaged in making “significant decisions” about individuals. Further, there is evidence to suggest the enthusiastic adoption of machine learning algorithms, along with worrying evidence of “historic bias in decision-making within these sectors” (Centre for Data Ethics and Innovation, 2019). While earlier practices of explicit discrimination on the basis of gender and minority status may not be displayed openly, they could have shaped the data held (Centre for Data Ethics and Innovation, 2019).

The report also observed that the financial services sector is unique insofar as it is highly regulated, has access to rich reserves of data, and a professional history of using modelling tools to aid in making complex decisions (Centre for Data Ethics and Innovation, 2019). This has placed the sector in an apt position to adopt

advanced data-driven technology. The financial services sector also constitutes a predominantly private sector use of algorithms, unlike say policing or local government. The interim suggestions of the report revolved around making algorithms more transparent and trustworthy, so as to ensure that algorithms help in improving decision making as opposed to worsening it. Moreover, the report recommended the use of more data and better algorithms to enhance the accuracy of risk predictions, and placing a mandatory transparency obligation on public sector organisations deploying algorithms that significantly affect individuals (Centre for Data Ethics and Innovation, 2019). This could mean that historically neglected populations that were unable to access credit earlier may secure better access in future.

In order to harness the potential of financial service automation in equalising finance and removing socio-economic barriers, it is essential that such automation adhere to fundamental principles of fairness, accountability, transparency and ethics in AI. Given the import of decisions taken by financial services firms on individual lives, it is worth scrutinising how such bias, historical or otherwise, may be eradicated to build a fair and ethical paradigm.

Regulation of fintech automation in India

The automation of financial services and the rise of robo-advisory applications is noticeable in India as well. While some automated applications may be designed to execute simple tasks using a few points of data, others may be engaged in complex tasks. An example of the latter would be an AI-driven robo-advisory application that is capable of analysing an investor's social media data to suggest a highly personalised portfolio.

Indian fintech companies appear keen to adopt robo-advisory due to benefits such as low operating costs, ease of scaling, and minimisation of human error and fraud. While the adoption is yet to reach mature levels, it could become widespread in the near future. Robo-advisory could bring potential benefits of financial inclusion, thus making it a particularly interesting issue for India.

As discussed above, there is no existing legislation regulating algorithms in India. On a related note, there is also no statutory right to seek explainability of algorithmic decision making in the financial sector or otherwise. The regulation of fintech companies and automation of services therein falls within the remit of the SEBI. It was clarified by the regulator in a 2016 Consultation Paper on Amendments/Clarifications to the SEBI (Investment Advisers) Regulations, 2013, that automation in financial services would be subject to the SEBI (Investment Advisers) Regulations, 2013, in addition to other relevant legislation.

The 2013 regulations lay down a framework for financial advisers acting in a fiduciary capacity towards clients. It also seeks to resolve instances of conflict of interest that may arise from the two roles played by the financial entity, both as an adviser and a distributor of financial products (SEBI, Memo on SEBI (Investment Advisers) Regulations, 2013, 2013). According to these regulations, 'investment advice' refers to advice on investing, purchase, selling or dealing in securities or investment products, and includes advice rendered through written, oral or other means of communication for the benefit of the client, and financial planning (SEBI, SEBI (Investment Advisers) Regulations, 2013).

The regulations set out various obligations for investment advisers such as obtaining a certificate of registration, furnishing of information, and meeting eligibility criteria. Investment advisers are expected to act in the best interests of the client in keeping with the relationship of trust. The adviser is obligated to ensure that all investments on which the investment advice is dispensed is in accordance with the risk profile of the client.

Other sectoral regulators have also attempted to regulate aspects of investment advisory relevant to their domains. For example, the RBI has released guidelines on investment advisory services offered by banks. The Pension Fund Regulatory Development Authority (PFRDA) has proposed regulations for regulating retirement advisers. However, in light of the absence of up-to-date regulations for automated investment advisory, the SEBI may also be considering imposing additional compliances for investment advisers using automated tools to provide

online investment advisory services (SEBI, Memo on SEBI (Investment Advisers) Regulations, 2013, 2013).

Further, the SPDI rules consider “financial information such as Bank account or credit card or debit card or other payment instrument details” to be sensitive personal data. This means that financial data is currently subject to heightened protection in terms of its collection, processing and disclosure. The Personal Data Protection Bill, 2019 has also treated financial data as sensitive personal data, and defined it as “any number or other personal data used to identify an account opened by, or card or payment instrument issued by a financial institution to a data principal or any personal data regarding the relationship between a financial institution and a data principal including financial status and credit history” (Personal Data Protection Bill, 2019).

The passage of the data protection bill¹ will have significant implications for the fintech sector and its adoption of automated tools. The DPA has the power to notify any data fiduciary or class of data fiduciary as a ‘significant data fiduciary’ under Clause 26 of the Bill. In arriving at such determination, the Authority will examine various factors like the sensitivity of the personal data processed, volume of personal data processed, turnover of the data fiduciary, risk of harm by processing, the use of new technologies for processing, and any other factor causing harm from such processing. It is likely that large financial services companies will accordingly be classified as significant data fiduciaries.

Further, if the Authority opines that processing done by any data fiduciary or class of data fiduciaries carries a risk of significant harm to any data principal, it can notify such data fiduciaries as deemed significant data fiduciaries (Personal Data Protection Bill, 2019). Under the Bill, “significant harm” refers to harm that has an aggravated effect considering the nature of personal data being processed (*ibid.*).

¹ The Personal Data Protection Bill, 2019 was introduced in the Lower House of the Indian Parliament on 11 December 2019 by the Minister of Electronics and Information Technology, Ravi Shankar Prasad. The draft legislation was then referred to a Joint Committee of Parliament (JPC) to deliberate, address concerns and submit a report on it. Once the JPC submits its report, the Bill will be reintroduced in the Parliament, which will then vote on it to make it an Act.

The significance of harm will be assessed on the basis of its impact, continuity, persistence or irreversibility. While ordinary harm includes financial loss, discriminatory treatment and denial or withdrawal of a service resulting from an evaluative decision about a data principal, significant harm would be of an aggravated nature (*ibid.*). Algorithmic bias could be responsible for harm or significant harm depending on the facts of the case.

Even though the Personal Data Protection Bill, 2019 does not explicitly regulate algorithmic processing in the financial services sector, or provide for explainability of automated decisions, its provisions are quite relevant to such processing. The provisions contained in the Bill to safeguard individuals from profiling-based harms will be especially relevant in the case of financial firms using automated tools to analyse risk profiles of clients. Further, it is likely that financial services firms will be subject to heightened obligations and compliances due to falling under the significant data fiduciary bracket.

It is worth noting that the DPA will look at the use of new technologies used for processing in its assessment. Financial services firms adopting automation tools like robo-advisory could thus attract the attention of the regulator. Given the gravity of decisions made by financial companies, they could potentially be classified as deemed significant data fiduciaries even if they do not meet the criteria contained in Clause 26 of the Bill. Moreover, it is possible that the DPA may expect financial companies deploying algorithms to make decisions, to be able to understand how a particular decision is being arrived at. It would be interesting to observe whether the DPA takes an individual notification approach, where particular financial services firms are notified as significant data fiduciaries on a case-to-case basis, or a class notification approach, where an entire category or sub-category of financial services firms are designated as significant data fiduciaries.

Recommendations for regulating financial service automation

It is understood that the present regulatory framework for addressing harms arising from automated processing is inadequate. In this regard, the financial services sector is no different. The overall need for fairness, transparency and

accountability that has been felt in algorithmic processing in general, is echoed in the case of fintech and robo-advisory applications. Further, considering the length and breadth of decisions made by such entities and the manner in which they affect lives of ordinary individuals, it is essential to seek explainability of automated decisions under their purview. Explainability is a practical means to achieve the goals of fairness, transparency and accountability. This is especially crucial in light of the fiduciary relationship between the financial adviser and the individual – and the potential for tremendous harm if the adviser is unable to discharge their duties appropriately. It should be an extension of the fiduciary relationship for the financial adviser to explain why and how a particular decision is being made for the individual.

Regulation of algorithmic processing in the financial services sector may be done by a proposed algorithmic accountability regulator (discussed below) in consultation with the SEBI. As the SEBI is the securities and commodities market regulator, it is well equipped to address concerns against financial firms. Further, the RBI would also have a say in such regulation with respect to banking entities. Other regulators like PFRDA and IRDAI may step in for issues pertaining to automation in the pension and insurance sectors. Such a framework would be buttressed by statutory guarantees to seek explainability of automated decision making in the financial sector.

This regulatory arrangement would achieve the advantage of having cross-sectoral scope, while ensuring informed inputs from sectoral regulators with respect to entities regulated by them. Inter-regulatory consultation of this nature is desirable, as the algorithmic accountability regulator may not possess the necessary domain knowledge or capacity to regulate the financial services sector effectively on its own (Andrews, 2017). Further, solutions to address automation in financial services may require a particularised approach instead of a one-size-fits-all approach across sectors.

As touched upon above, two areas of particular significance that would require regulatory attention are algorithmic bias and the right to explainable decisions. With regard to the first area, the CDEI in the UK has warned that reliance on

historical data may lead to biased outcomes in the present day. For instance, past lending data that captures which borrowers were good or bad credit risks could be masking credit invisible individuals belonging to marginalised groups. This may create a vicious circle where credit invisible individuals are denied credit because there is a lack of supporting data, and they continue to remain credit invisible. Feeding in biased data of this nature could manifest in biased algorithmic decisions on creditworthiness that are potentially racist, classist, sexist or communal. In light of this, it is essential to prevent algorithms from persisting or amplifying historical biases, or introducing new biases into the decision-making matrix. Fair and unbiased decisions are good for all stakeholders involved — individuals, businesses and society (Centre for Data Ethics and Innovation, 2019). Thus, the need for trustworthy algorithms should not be underestimated.

With regard to the second area, explainability of decisions in financial services is required due to the manner in which they can adversely affect individuals. It is therefore essential for financial services firms deploying automated tools to be aware of the rationale of their algorithms, and be able to explain which factors were given weight and how the decision was made (Krishnan, Deo, & Sontakke, 2020). Explainability of decisions in financial services must be accounted for in any potential statutory regime carving out the right to explainability, due to the financial sector's unique position to impact individuals.

Considering the rapid adoption of advanced technological tools by firms, the cross-sectoral regulator and sectoral regulators may consider the use of regulatory technology or RegTech tools (Krishnan, Deo, & Sontakke, 2020). Such tools could enable regulators to stay on top of developments by tapping into technological solutions to regulatory limitations. RegTech tools could enable regulators to ensure early detection of fraud, bias and other violations.

Further, the proposed algorithmic accountability regulator should consider permitting financial service firms to participate in regulatory sandboxes to demonstrate the use of new automation techniques. Such sandboxing activities could be supervised by the algorithmic accountability regulator along with the

relevant sectoral regulator like SEBI, RBI or IRDAI. This could work in a seamless manner, as these sectoral regulators have already put in place their own regulatory sandbox regimes.

The financial services sector in India should suitably prepare itself for the upcoming legislative overhaul in the form of the Personal Data Protection Bill, 2019. Further, while the data protection law will set the stage for rights of data principals and duties of data fiduciaries, it may not be entirely adequate to address the harms arising from automation in the financial sector. It is therefore recommended that the data protection law is used as a base to build on further regulatory paradigms that can ensure fairness and transparency in finance. Chief among potential future efforts to achieve fairness and transparency in finance would be seeking explainability from automated decisions arising in the financial services sector. In order to make machine-made decisions in the financial sector accountable, it is necessary to examine the challenges of implementing a right to explainability in India.

Environmental factors in implementing the right to explainability in India

As discussed above, the global debate on algorithmic accountability has not converged on a universal benchmark yet. Within the four corners of this debate, however, the right to explainability has emerged as a significant demand for securing individual autonomy. In India, individuals have not been given explicit protection against particular harms arising from automated decision making or profiling in the Indian Personal Data Protection Bill, 2019. This is primarily because the right to explainability of automated decisions has not found place in the Bill yet.

There are broadly three reasons why despite the excitement surrounding it, the right to explainability has not yet featured in mainstream legal frameworks on data protection and privacy. *First*, the right to explainability is not sufficient in and of itself. In order to be effective in securing individual autonomy over their personal data, the right requires the support of a network of other data protection rights and obligations. These supporting rights and obligations include the

obligations for fair and transparent processing, purpose limitation and collection limitation; and the rights to confirmation, access, correction and erasure.

Second, there is a lack of consensus on whether providing meaningful explanations about the outcomes of machine learning decisions is even practically possible. Two key attributes of algorithmic decision making have been identified as its opaque and automated nature (Zarsky, 2015). While arriving at a decision, an algorithm adopts non-transparent means and the data analysis is done automatically. Sceptics believe that since deep learning is a black box by its very nature, it is not possible to simply glance inside a deep neural network and find out how it works (Knight, 2017). As mentioned earlier in the paper, Knight has warned that providing individuals with explanations even for simple systems may be impossible (*ibid.*). Such scepticism and ambiguity about the workings of algorithms may have deterred policymakers and legislators and made them wary of prematurely including a right to explanation within laws.

Third, a legal requirement to implement explainability may have a significant impact on regulated businesses. The impact could be felt in the form of compliance costs in providing requested explanations to individuals, as well as the potential conflict of such a right with their trade secrets and intellectual property (Wachter, Mittelstadt, & Floridi, 2017). This may disproportionately impact Micro, Small and Medium Enterprises (MSMEs) and start-ups, vis-à-vis large corporations with greater wherewithal. In fact, it is not yet clear how much implementing a right to explainability could cost the economy in the short term and long term.

Nevertheless, in keeping with the global debate, it is useful to understand which considerations are relevant in formulating and operationalising a right to explainability. Such a right should be inserted to be part of a larger data protection and privacy legislation in the relevant national jurisdiction. For example, the right would be part of the GDPR in the EU, and the PDP Bill, 2019 in India. It could also form a part of separate legislation on algorithmic accountability.

It is particularly crucial for national governments to formulate the right to explainability in the context of their nation's unique circumstances. An inability to do so may set up the proposed legal and regulatory measures for failure, due to their unsuitable character. For instance, it is evident from Indian policymaking over the last few years that the Indian government wishes to boost the digital economy and encourage domestic MSMEs. Keeping this in mind, the government may decide to adopt a graded approach in operationalising a right to explainability – where the distribution of compliance burden differs on the basis of the size of the concerned entity.

The manner in which the provision on the right to explainability is drafted is significant. As pointed out by some scholars, the presence of ambiguous legal drafting of the right could render it ineffective (Wachter, Mittelstadt, & Floridi, 2017). It is thus essential that the right is drafted in a clear, concise and simple manner, making its intent clear to both individuals and entities in order to suitably inform their respective conducts. This is perhaps the single most relevant factor in realising a meaningful right to explanation. Latent ambiguities have the potential to derail the legislative intent and render the right ineffective. For example, the drafting of the right to explanation in the EU GDPR has led to considerable confusion, with scholars being divided on whether such a right has even been guaranteed (Andrew D Selbst, 2017). This is, *inter alia*, due to the absence of a single neat statutory provision on the right to explanation, and the use of vague language such as “meaningful information about the logic involved” (*ibid.*).

Moreover, considering that such a right may have a tectonic shift on the digital economy, its formulation should be done after conducting widespread stakeholder consultations. This would not only aid the government in assessing the potential costs of implementing the right, but also enable a conversation between the state and industry on how the latter can suitably tweak their internal systems to comply. Having said that, the right to explainability is an individual-centric right that is meant to provide control and autonomy to the individual or data principal. Any formulation of the right must reflect this fundamental truth.

To further nourish the right to explanation, a range of broader measures may be additionally contemplated. For instance, Kumar (2019) has recommended an algorithm transparency bill with a range of safeguards such as external audits for bias that are made available to the public, diversification of workplaces, enhancing algorithmic literacy among users and encouraging research on algorithmic techniques for reducing human bias (Kumar, 2019). Further, he has recommended the updating of extant laws to suit the digital domain (Kumar, 2019). Chatterjee and Ravindran (2019) have suggested that potential solutions could be both legal and organisational. While legal solutions include passing a strong personal data protection law armed with the right to logic of automated decisions and a general anti-discrimination law, organisational measures focus on transparency and redressal mechanisms provided by organisations themselves (Chatterjee & Ravindran, 2019).

Crawford and Schultz (2019) have argued for holding vendors supplying AI-based decision-making tools to the government liable (Crawford & Schutlz, 2019). Kearns and Roth (2020) have suggested arming tech regulators to encourage deeper investigations. Regulators could discover algorithmic misbehaviour in a controlled and confidential manner by accessing the underlying code (Kearns & Roth, 2020). The authors have argued that regulators and their powers must evolve sufficiently in order to tackle the new challenges of technology, by, for example, gathering the capacity to conduct algorithmic audits at scale. In doing so, they can look to the theory of ethical algorithm design for guidance (*ibid.*).

Furthermore, a range of self-regulatory measures have been suggested by various authors such as operators of algorithm regularly auditing for bias, working with diverse teams with diverse expertise, and increasing human involvement in the design and scrutiny of algorithms (Lee, Resnick and Barton 2019). All of the above proposals are worth considering in their own right, and should inform future policymaking on algorithmic accountability keeping in mind jurisdictional needs.

Section VI: Situating the right to explainability in the Indian context

While the global debate on the right to explainability can inform and shape India's policy, the ultimate formulation of the right must be determined by the country's unique circumstances. In this regard, it is worth cautioning against the unthinking transplantation of legislative and policy frameworks from other jurisdictions. Scholars have warned against adopting a standardised cookie-cutter approach to AI regulation (Hickok & Basu, 2018). Merely because a particular formulation has worked in another country does not mean that it will also work in India, especially in the relatively uncharted waters of the right to explainability.

The following three factors may serve to distinguish India from other jurisdictions, and should consequently inform policymaking. *First*, the size and population density within the Indian territory must be accounted for. India is the seventh largest country and the most populous democracy in the world. The unwieldy size of the country cautions against the setting up of overtly centralised regulatory regimes. Instead, it necessitates the setting up of devolved and decentralised enforcement mechanisms that can reach all parts of the country and provide a redressal forum to citizens residing even in remote areas. However, existing frameworks generally leave much to be desired, with most leaning towards overt centralisation and lack of local presence.

Second, not only is India spatially expansive and population-dense, it is also characterised by remarkable linguistic heterogeneity. Article 343 of the Constitution of India accords the status of 'official language in the Union' to the Hindi language in Devanagari script. Further, the Official Languages Act, 1963 mandates the publication of the law only in English and Hindi.

Nevertheless, this is not entirely representative of the linguistic diversity in the country. It leads to the exclusion of a vast majority of the population who are non-English speaking and non-Hindi speaking, and thus unable to participate in public consultations, become aware of existing law and policy and enforce the rights they enjoy thereunder. This is despite the fact that Schedule VIII of the Constitution recognises twenty-two languages including Gujarati, Bengali and Tamil. Linguistic heterogeneity is particularly relevant in envisaging a right to

explainability. For a right that is intended to empower the ordinary individual in her navigation of the digital wild, an inability to make the law accessible to all would be a significant concern.

Third, since India is characterised by vast income and wealth inequality, large swathes of the population have traditionally not enjoyed access to the internet. As per a report by the Internet & Mobile Association of India and Nielsen in 2019, less than 50 percent of the Indian population is active on the internet (India I. a., 2019). Matters get grimmer when it comes to usage by women in India, who account for half the number of total male users. Rural users enjoy an even slimmer portion of the pie — at 27 percent, despite accounting for nearly 65 percent of the total population of India.

Moreover, even where access has been resolved, quality of internet continues to remain an issue. Scholars have recognised the difficulty in creating a single umbrella regulatory framework for the use of AI in a country with diverse socio-economic demographics like India (APRU, 2020). Access and quality issues erect a digital divide that only serves to exacerbate existing socio-economic inequalities. It has thus been argued that the internet should be treated as a priority public utility (Srivastava, 2020).

It is essential for prospective regulation on AI and the right to explainability, to acknowledge the above factors during formulation of policy. Giving due consideration to India's size and population density, linguistic heterogeneity and income inequality would mean taking a step back from ideas about algorithmic regulation developed in a Western context. In fact, doing so could ease the process of striking a balance between competing interests such as privacy and innovation, and algorithmic performance and explainability.

Furthermore, it is also useful to understand the conditions in which regulations in India are embedded. For example, the NITI Aayog's National Strategy for AI attempted to curate its contents keeping in mind ground realities unique to the Indian situation. It has earmarked priority areas of focus where AI intervention may be explored, such as healthcare, agriculture, education, smart cities and

transportation. It has also recommended incentivising research in AI, skilling for the AI age and accelerating adoption of AI in the value chain (Aayog, 2018). From a regulatory perspective, it has suggested self-regulation by stakeholders and benchmarking data protection laws to international standards like EU GDPR (Aayog, 2018). However, it is not clear whether such a regulatory approach would entirely meet the challenges posed by the three factors discussed above.

The Srikrishna Committee of Experts has been acutely aware of India's position as an emerging digital economy. Curiously, its attempt to promote innovation for digital enterprises may have prevented the inclusion of a right to explainability in the draft Personal Data Protection Bill, 2018 floated by the Committee (Srikrishna, *A Free and Fair Digital Economy: Protecting Privacy, Empowering Indians*, 2018). We will explore below what such regulatory constraints in the Indian context could look like.

First, the culture around privacy and data protection in India is still not adequately mature. This is primarily due to the belated development of law and policy on privacy, as evidenced by the absence of a personal data protection law till date. In fact, the issue of whether the right to privacy is a fundamental right was not settled till as late as 2017 in the *Puttaswamy* case. Coupled with the rapid pace of technological evolution — and new kinds of technology solutions springing up every day — the Indian government's regulatory stance towards technology has been somewhat erratic.

The official regulatory approach has swayed between two extremes, with robust technocratic optimism on one hand, and outright dismissal of new technology on the other. The government's faith in the Aadhaar project is an example of the former, while the recent ban on cryptocurrencies imposed by the RBI is an example of the latter.² In the former instance, the government has insisted on

² Aadhaar is a unique identification scheme in India that is based on biometric and demographic data. It is overseen by the Unique Identification Authority of India, a statutory authority created by the Aadhaar (Targeted Delivery of Financial and other Subsidies, benefits and services) Act, 2016.

making Aadhaar an inevitable part of the lives of citizens despite evidence of its grave dangers of perpetuating exclusion of already marginalised groups, infringing upon privacy, and legitimising a surveillance state. In the latter instance, the RBI has refused to objectively evaluate the merits of permitting cryptocurrencies despite not having sufficient evidence to ban it.

Second, regulators in India often suffer from capacity constraints and lack of resources necessary to enforce regulation. This is particularly concerning in the case of the right to explainability, because the realisation of the right may be resource-intensive. Since the government itself is the largest data fiduciary, it would need to ensure its own compliance with the mandate of the right. It would also need to enforce the law and ensure that adequate grievance redressal mechanisms are put in place.

Further, regulators may also lack the requisite expertise to regulate emerging issues of technology. Most regulators are staffed by career bureaucrats and politicians who may not possess the domain knowledge required in tech policy issues. This underlines the need to have experts on board, who can offer technical insight and enhance the internal tech literacy of the regulator.

Third, in many ways, India is a young economy that is playing catch-up since liberalisation to more dominant global forces. The thriving tech industry driving datafication and digitisation across the board is an important component of the Indian economy. In light of this, policymakers should be aware that heavy-handed regulation, or regulatory uncertainty, may quell innovation due to compliance fatigue or fear of liability (APRU, 2020). It appears that the Srikrishna Committee of Experts was aware of this aspect. It chose to capture the need to promote the digital economy at the very outset in the title of its 2018 report, aptly named 'A Free and Fair Digital Economy: Protecting Privacy, Empowering Indians'. Perhaps for the same reason, it declined to include a right to explainability in the draft Personal Data Protection Bill, 2018. Thus, regulators must walk the precarious tightrope of promoting innovation while simultaneously safeguarding individual rights and autonomy.

Fourth, the federal structure of the government has encouraged states to attempt regulation of AI in their own ways, especially in the absence of clear central regulation on the matter. This may pose a challenge for federal regulators in future, who seek to formulate an overarching framework on AI regulation and algorithmic processing of data.

For example, the Tamil Nadu government has recently released the Safe and Ethical Artificial Intelligence Policy 2020, and the setting up of the Centre of Excellence in Emerging Technologies. The initiative is intended to solve key governance issues with the help of AI, IoT, blockchain and augmented reality/virtual reality. The policy document provides a roadmap for safe and ethical deployment of AI, covering transparency, audit, misuse protection, digital divide and data deficit. A state-wise approach makes sense because local governments are best equipped to understand indicators such as internet penetration, financial literacy, and tech literacy within their borders.

However, as noted above, a fragmented mosaic of AI frameworks of this nature may pose a harmonisation problem for central regulators in the future, and impede the crystallisation of concrete norms at the central level. Existence of various state-wise norms may also pose a challenge to industry, as players would need to conform to varying regulations depending on the state they are operating in. This could lead to ambiguity and lack of regulatory certainty among industry players, thus disadvantaging the overall progress of the digital economy.

Section VII: Adjusting the FATE framework to the Indian context

As discussed earlier in the paper (and in more detail in Appendix A), FATE is a succinct framework for studying algorithmic decision making, and has witnessed a rise in popularity over the last few years. It provides a concise analytical tool to assess how algorithms fare on key parameters. The principles of fairness, accountability, transparency and ethics in AI serve to ensure that AI models function within acceptable boundaries. In other words, it seems that algorithms are making unbiased decisions, responsibility is being assigned for decisions

made by algorithms, organisations are open with their end users, and the ethical dimensions of building automated systems are kept in mind. Moreover, the ambit of FATE has been expanding, with the 'E' being additionally interpreted as explainability, and 'S' added to signify safety and security (Wing, 2018).

Despite the popularity of the FATE framework, it is worth remembering that it has been developed in a specific Western context suited to the regulatory and socio-economic realities of those jurisdictions. Akin to the development of algorithms themselves, the lack of representation may limit the utility of the framework in other jurisdictions and contexts. This is because other jurisdictions may differ significantly on various parameters vis-a-vis Western countries. It is thus necessary to examine FATE in the Indian context, and identify any adjustments that may be required.

The discussion in the previous chapter has indicated some factors that may set India apart from other jurisdictions and should inform policy formulation. Broadly speaking, these are size and population density, linguistic heterogeneity, and income and wealth inequality. In addition to these, the earlier discussion highlighted some regulatory constraints in the Indian context. The interaction of these factors creates an elaborate web that policymakers must account for.

At the outset, it must be acknowledged that a widespread adoption of FATE could usher in significant benefits. It is worth noting that the extant legal and regulatory framework on algorithmic processing is currently nebulous in India. As a result, several critical questions are either left unaddressed, or with plenty of room for interpretation. This could prove to be detrimental to ethical AI in the long run. Characterised by an absence of a privacy and data protection legislation, and a culture around privacy that has not yet reached maturity, FATE could bring theoretical underpinnings and conceptual cohesion. It could potentially provide much-needed awareness and guidance to developers of algorithms and entities deploying algorithms, by instructing them on the broad parameters they should be attentive to. In light of this, it is recommended that Indian policymakers curate the FATE framework in the Indian image. This could catapult FATE as an effective and customised solution to address algorithmic bias.

For instance, the principle of fairness should be informed by constitutional law jurisprudence of the Supreme Court and the high courts of India. Jurisprudence on the fundamental right to equality contained under Article 14 of the Constitution of India, 1950 has explicated on the concepts of 'fairness', 'arbitrariness' and 'reasonable classification' in various landmark cases. While such jurisprudence was evolved for assessing decisions made by human beings, now they can be suitably modified to extend to machine-made decisions. The resultant meaning of 'fairness' should account for size and population density, linguistic heterogeneity, and income and wealth inequality. This would mean that the outcomes of algorithmic decisions should be fair for all Indians – irrespective of "religion, race, caste, sex or place of birth" (Article 14, 1950).

Similarly, the principle of accountability should operate in a manner that individuals are able to repose their trust in entities relying on automated processing. Accountability is especially important in a country characterised by wealth disparities. Over the past decade, Indians have observed, been subjected to, and suffered the excesses of a top-down technology powered digital identity behemoth called Aadhaar. The working of the Aadhaar project has resulted in pervasive injustice, exclusion and perpetuation of existing socio-economic inequalities. If the government expects to make progress from this murky technological past, it must make automated decision-making sufficiently accountable. Such a move will promote trust and confidence, as well as nourish the privacy culture in the country.

The principle of transparency should be interpreted expansively, in order to ensure that not only is there openness around the final decision, but also around the process of arriving at the decision and the trade-offs being made by the entity. Users should know that the decision affecting them has been made by a machine, whether there was or was not any human intervention, what was the process of arriving at the decision, and the sort of trade-offs deemed acceptable by the concerned entity. For example, if my credit application request was processed by an algorithm, I deserve to know the upside and downside of such automation. This could act as a check on entities in balancing efficiency and fairness.

Ethics may be viewed as an umbrella term capturing all of the above principles. Since the ambit of FATE has been expanding, with 'E' standing for explainability, it is worth reviewing its application. The implementation of explainability would simultaneously help realise fairness, accountability, transparency and ethics. In a sense, it may be said that explainability is the practical strategy for implementing the above ideals. Explainability may be operationalised as a legal right, contained either in the data protection law or a separate law on algorithmic accountability. However, it must be mindful of the vast heterogeneity and disparity in India. A meaningful version of the right would strive to ensure that each individual, no matter their income, language, or location, is able to obtain an explanation of automated decision-making affecting them, in an easy and accessible manner.

Nevertheless, implementing the FATE framework may face some roadblocks. Some of these are in the nature of regulatory constraints, that have been discussed in the previous chapter. To summarise, these constraints are the lack of maturity in privacy culture, lack of resources and state capacity, need to promote enterprise, and federal structure of government. These constraints may have significant implications on the feasibility of implementing the FATE framework from the vantage points of regulators and enterprises. For regulators, the lack of resources and state capacity may make the real-time supervision of entities' compliance with FATE obligations an uphill task, given the volume of automated decisions being made. Further, policymakers and regulators would need to strike a balance between promoting the digital economy and its constituent enterprises, and securing individual autonomy.

For entities themselves, complying with FATE obligations may increase costs, become a time-consuming affair, and add to a list of already lengthy compliances. For entities that do not have a data protection officer on their payroll till now, it may become an uncomfortable transition. This could be an aggravated blow to MSMEs, which contribute substantially to India's GDP and are a source of income for more than 10 percent of the population (Sharma, 2020). More significantly, the debate on the principles of FATE remains foggy to some degree. For instance, there is no unanimous conclusion in support of whether meaningful explainability

is even possible. Such factors may demand entities to re-orient their existing business models, and go back to the drawing board to arrive at an ethical AI compliant paradigm.

Section VIII: Empirical Research

This paper looks at three key categories of stakeholders – regulators, users and private sector who utilise and manage ML and AI-based applications. Such applications have a wide and deep sectoral reach that spans a range of fields such as healthcare, retail, agriculture, manufacturing, transport, energy, smart cities or urban development, education and skilling, telecom, and of course, the software and IT industry itself.

Regulators form an important part of the AI landscape. They, simply put, oversee a certain policy area and introduce checks and balances to enable transparent and fair processes. Therefore, the key function of the regulator is to apply regulations that ensure fairness and stability of a system. But, in a rapidly evolving technological era, the task of a regulator has become critical as they are required to constantly update and modify these regulations. This leads to enforcement challenges for the regulators and compliance challenges for entities that keep track of these developments. This, in turn, has increased the need for AI and analytics in regulatory compliance to ease the burden on all stakeholders involved.³ For the users, AI algorithms offer far better speed and reliability at a much lower cost compared to human-interventions. Users often encounter AI technology and solutions in areas such as e-commerce, workplace communication, human resource management, healthcare, cybersecurity, logistics and supply chain, sports betting, streamlined manufacturing, public administration and services among others. This is also witnessed in daily use devices like smartphones, digital personal voice assistants like Siri; web search,

³ <https://www.forbes.com/sites/cognitiveworld/2019/07/22/why-regulatory-compliance-can-be-complicated-and-how-ai-can-simplify-it/?sh=5cf787f6377e>

machine translations etc. For those who belong to this category, there are two types of AI use-cases – software and embodied. ‘Software’ cases include virtual assistants, image analysis software, search engines, speech and face recognition systems while ‘Embodied’ cases comprise robots, autonomous cars, drones, Internet of Things.⁴ For the private sector, the use of AI is potentially transformative. Due to increasing digitalisation, there is a need to continuously generate, process and analyse data. In the private sector, AI has been the focus of research for more than 30 years. It is therefore no surprise that the embrace of AI and ML-led systems by the private sector has enabled the creation of new business opportunities, as evidenced by a PwC study that found out that 62 percent of large companies are already utilising AI. The much-touted Industry 4.0 will be fuelled by private sector use applications that are built on state-of-the-art end-to-end IT infrastructure that can withstand cybersecurity risks.

In terms of responsible AI, regulators have generally identified 6 principles — fairness and unbiasedness, security and safety, privacy, inclusiveness, transparency and accountability that should underpin any national or enterprise-led strategy. In the Indian context, experts have pointed to five ways of ensuring responsible AI in the regulatory frameworks: a) adopting a system of agile governance, b) creating a developmental sandbox (which will test an algorithm for its output), c) establishing a set of ethical standard and principles, d) creating an index that measures the balance and e) being responsible by design.⁵ Others have identified the need for a uniform definition of AI for regulators (on the basis of law) and to setup an independent body to test the unbiasedness of any AI algorithm. Regulators, in the course of developing responsible AI, must regulate not only the data but also the performance metric by which an AI is judged. For users, algorithmic fairness is necessary to understand the second and third-order effects of AI use applications. This is borne out by an Oxford study in which 82

⁴ <https://www.europarl.europa.eu/news/en/headlines/society/20200827STO85804/what-is-artificial-intelligence-and-how-is-it-used>

⁵ <https://indiaai.gov.in/article/role-of-regulations-for-responsible-ai>

percent of respondents believe that AI should be carefully managed.⁶ Ethics matter to end users who trade their data for efficient solutions. The data science and AI community realizes it has the power to advocate for how they would like to do the work. Given that it is key to retain consumer trust & satisfaction, private sector organisations are keen to develop ethical, and more recently, FAT (Fairness, Accountability and Transparency) AI in their interactive platforms and processes. However, studies have shown that nearly 90 percent of companies (85 percent in India) have encountered challenges in ethical AI.⁷ The private sector must build awareness, ensure diversity within teams tasked to build AI systems and develop governance structures etc. In the NITI Aayog's June 2018 paper titled National Strategy for Artificial Intelligence, contributions were made by private sector companies like Wadhvani AI, NVIDIA, Intel, IBM, NASSCOM, McKinsey, Accenture, MIT Media Labs etc, highlighting the pressing need to align organisational goals with AI ethics and principles.⁸

The concepts of robustness and explainability of AI systems have emerged as key elements for a future regulation of this technology. Regulators are exploring the known vulnerabilities of AI systems, and the technical solutions that have been proposed in the scientific community to address them. One of the ways is through the promotion of transparency systems in sensitive systems through the implementation of explainability-by-design approaches in AI components that would provide guarantee of the respect of the fundamental rights.⁹ As more regulators become aware of the issues around 'Explainability', 'Provability' etc, concepts such as 'Differential Privacy', by implementing 'Federated Learning' wherein data trusts are developed for easy and secure sharing of data without compromising any sensitive personal data or information, achieves greater

⁶ <https://www.bcg.com/publications/2020/six-steps-for-socially-responsible-artificial-intelligence>

⁷ <https://indiaai.gov.in/article/its-2020-are-we-still-questioning-the-importance-of-responsible-ai>

⁸ https://niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf

⁹ <https://ec.europa.eu/jrc/en/publication/robustness-and-explainability-artificial-intelligence>

salience.¹⁰ For users of AI technology, the upward trend in sophistication and complexity (decision making by algorithmic black box) calls for greater explainability. In order to retain stakeholder trust, it is necessary to know the rationale of how the algorithm arrived at its recommendation or decision. Even businesses will be better positioned if they adopt good practices around accountability and ethics by building interpretability into AI systems. As efficiency is crucial to the fortunes of a private sector business in the AI space, it is equally vital to maximise performance by identifying potential weaknesses. Explainability is a powerful tool for detecting flaws in the model & biases in data and can help verify predictions, improve models and gain insights into the problem at hand. For existing AI implementations, the way to introduce AI explainability is through gap analysis and the private sector is increasingly mindful of this evolving feature.

In terms of the trade-off between explainability and accuracy, it is important for the human-computer interface to translate the model to an understandable representation for the end users. By and large, there are three steps to follow for a typical AI system: a) Explain the intent behind how the system affects the concerned parties, b) Explain the data sources you use and how you audit outcomes, and c) Explain how inputs in a model lead to outputs.¹¹ However, the drawback is that, since they are simple, explainable models don't work very well. There are doubts over whether it can accurately fulfil the task it was designed for. On the flip side, the lack of interpretability leads to reduced transparency and accountability of predictive models which can have (and has already had) severe consequences; there have been cases of people incorrectly denied parole, poor bail decisions leading to the release of dangerous criminals, ML-based pollution models stating that highly polluted air was safe to breathe, and generally poor use of limited valuable resources in criminal justice, medicine, energy reliability,

¹⁰ <https://www.mondaq.com/india/privacy-protection/1015476/self-regulation-in-artificial-intelligence-an-indian-perspective>

¹¹ <https://www.kdnuggets.com/2019/01/explainable-ai.html>

finance, and in other domains. Therefore, explaining the decisions made by ML models is the need of the hour and there has been a steady increase in demand for transparency and accountability of such complex systems, there has been a recent explosion of work on “Explainable ML”

AI model performance is estimated in terms of its accuracy to predict the occurrence of an event on unseen data. A more accurate model is seen as a more valuable model. Despite the unimaginable computing power and cost reductions, until now most developers emphasize more on incrementally improving the performance of AI systems according to a narrowly defined set of parameters and not on how the algorithms are achieving the requisite success. Model interpretability can be summarised as providing insight into the relationship between the inputs and the output. An interpreted model can answer questions as to why the independent features predict the dependent attribute. Issues arise because as model accuracy increases so does model complexity, at the cost of interpretability.

Therefore, the trade-off decision to be made should depend on the application field of the algorithm and the end-user to whom it is accountable. When dealing with technical users with high level of sophistication and trust level, accurate models are preferred over high explainability as performance is very important. However, with respect to use applications for the lay user, as it is usually in the regulated space such as banking, insurance, healthcare etc, AI-based firms are prone to legal and ethical requirements that limit the use of black box models. In such a scenario, users are better served by simple XAI although the algorithm might be less accurate. Ultimately, the success of AI models is due to the machine’s own internal representations which are more complicated than manually generated features leading to its inexplicable nature. The best bet for users, regulators and the private sector would be to build and manage interpretable models that can work in tandem with human experts and their specific area of expertise.

Research Methodology

This paper aims to explore the unanswered questions regarding policy concerns due to algorithmic decision making, specifically in the context of the fintech and robo-advisory space in India. This paper aims to report the true extent of the pervasion of ADS in the financial services sector. This is important, as given the extent of pervasion, appropriate regulations can be tailored for the sector. Further, given the maturity of the responsible AI movement in India, it aims on identifying the optimal mode of regulation. The international comparison of algorithmic regulation and governance across jurisdictions, which was undertaken in Section III will serve as the base study to analyse. The paper also aims to provide insight specific to India by accounting for the unique constraints operating in the Indian context. This will take into account the effort of the Securities and Exchange Board of India, Reserve Bank of India and the Pension Fund Regulatory Development Authority. Finally, this paper aims to establish a greater understanding of the trade-off between algorithmic explainability and algorithmic performance. The case for implementing a right to explainability in India has been previously specified (Section VI). There is a need to understand the factors which distinguish India from other jurisdictions and potential regulatory constraints at play in India.

The aim of this research will be to understand the perspectives of the sampled individuals on the adoption of algorithmic processing India, as well as the concomitant issues surrounding such adoption. In doing so, questions will be posited by the researcher, hinging on a set of key themes. These will be as follows: a) pervasion of ADS in the financial services sector, b) maturity of the responsible AI movement in India c) appropriate modes of regulation, d) unique constraints operating in the Indian context, and e) trade-off between algorithmic explainability and algorithmic performance.

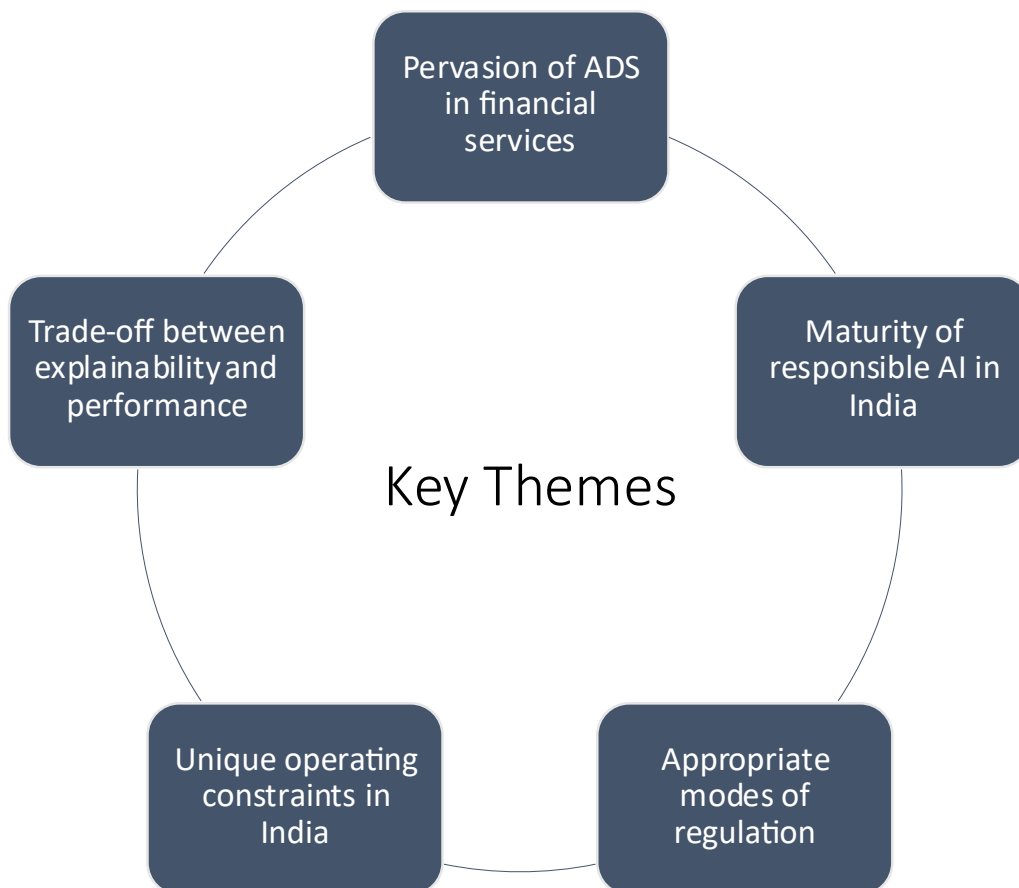


Figure 1. 2 Key themes in the research methodology of this paper

The discussion contained in the previous chapters make it clear that direct transplantation of ideas and frameworks from other jurisdictions may have limited utility. This is due to the unique realities of the Indian jurisdiction. The methodology to be adopted has been decided keeping this in mind. A qualitative methodology has been adopted to shed further light on appropriate regulation of algorithmic decision-making in India. Such a qualitative methodology is essentially more fluid and flexible than its quantitative counterpart, making it particularly suitable due to the room it will provide for a voyage of discovery (Bryman, 1984).

Semi-structured interviews will be conducted with a select group of participants based on the abovementioned themes. Such interviews will enable the understanding of the issue of algorithmic governance from the point of view of the concerned actor. A semi-structured interview technique entails that the

interviews could be a mix of structured and unstructured questions. While it would provide direction and purpose, it would also leave room for unanticipated findings and the possibility of altering research plans if needed.

To analyse the data collected, a thematic analysis will be done. Thematic analysis is a method to identify, interpret and analyse patterns of themes within qualitative data. It would enable the identification of crucial insights from the interviewees on ADS infiltration in the financial services sector, in a manner that is structured yet flexible. Therefore, this would constitute a rigorous way to attain qualitative insights that we are seeking.

The observations provided by the interviewees would also enable the weighing of two competing concerns – i.e., whether to prioritise explainability or algorithmic performance. Based on whether the interviewee is situated in the private sector, academia or the government, it is possible to glean insights on why one of the competing interests is considered superior over the other.

Hypotheses

Interview-based evidence is the source of empirical information in this project. These interviews will help evaluate the hypotheses for each of the undertaken themes. Regarding the pervasion on ADS in the financial services sector, it is hypothesised that the implementation of robust regulations to create a transparent and technologically-advanced financial services sector is unlikely. This is possibly due to private players distorting the market.

With regards to the maturity of responsible AI in India, it is likely since the need for this movement in India is a must; India will catch up to its Western counterparts. The importance AI gains among relevant stakeholders will only further this movement.

For the appropriate mode of regulation, it is hypothesised that the potentially significant impacts AI will have in India's future will be recognised by its population. However, it is unlikely that a purely self-regulatory model will be favoured due to doubts regarding efficacy of the implementation of the model.

While considering the unique constraints operating in the Indian context, it is likely that since ADS is a new topic in India, the stakeholders may not have considered this up until very recently. Still, it is expected that the interviewees would be aware of the nascence of the responsible AI agenda in India.

Regarding the trade-off between algorithmic explainability and algorithmic performance, an inverse relationship is expected. This is since there is a need to prioritise the value of one over the other. There will definitely be a challenge in making AI-driven data decision making to be made explainable, especially while aiming to have transparency around these decisions.

Hypotheses

- 1** **Theme:** Pervasion of ADS in the financial services sector
Hypothesis: Implementation of robust regulations to create a transparent and technologically-advanced financial services sector is unlikely at current stage
- 2** **Theme:** Maturity of the responsible AI movement in India
Hypothesis: Likely, since need for this movement in India is a must; India will catch up to the West
- 3** **Theme:** Appropriate modes of regulation
Hypothesis: Unlikely that a purely self-regulatory model will be favoured due to doubts regarding efficacy of implementation
- 4** **Theme:** Unique constraints operating in the Indian context
Hypothesis: Likely that ADS is still a new topic in India, but interviewees will be aware of its early stage
- 5** **Theme:** Trade-off between algorithmic explainability and algorithmic performance
Hypothesis: Inverse relation between explainability and performance is expected

Figure 1. 3 Summary of hypotheses

Interviews

Interviews are essential methods required to understand contemporary actions and their outcomes. The causal mechanisms interviews are able to identify are not evident in their quantitative counterparts. They can serve as a central source of data or even generate data that will be used in future statistical analyses. Despite the typical sample size being smaller than that of surveys, interviews

generate a deeper set of responses. Depending on the answer received, the interviewer has the option to ask follow-up questions to probe deeper into the attitudes and actions of the respondent. These are especially valuable in the case of contradictory views in the response. None of this is possible in surveys due to length and cost limitations. But most importantly, with interviews, the interviewer has the metadata at his/ her disposal (Mosley, 2013).

Even a single interview can generate more points of inferential leverage. It can provide information regarding the attitude held or actions taken by others. The interviewer knows how the respondent behaved and if they showed signs of hesitation for certain questions. This will facilitate a more accurate interpretation of their data as compared to quantitative indicators (Mosley, 2013).

It is also important to consider the 'interviewer effects' on the respondent. This is especially significant in terms of gender as it often affects the interviewee's response (non-response) to a question (Mosley, 2013). Other challenges considered include ethics and sampling methods.

Ethics

It is common practice to ensure that the respondents are clearly explained the benefits and risks of their participation in the study. Naturally, the risks will be minimized and be within reason to be ethically acceptable. Typically, studies in social science do not pose a high risk for the respondents, but disclosure from interviewers regarding how the data may have negative social or professional consequences is common practice. Informed consent procedures are a must and, in most cases, the confidentiality of the participants in terms of identity and participation is also required (Mosley, 2013).

Explicit consent has been sought from the interviewees for the purpose of this research paper. For reasons of personal privacy and organisational conflicts, the identity of the interviewees will be kept confidential (Mosley, 2013).

Sampling

The next challenge is with regards to selecting the appropriate sample. Based on semi-structured interviews, this paper employs the purposive sampling method

which, according to Springer Link, is the 'intentional selection of informants based on their ability to elucidate a specific theme, concept, or phenomenon.'¹² It involves an iterative process of selecting research subjects rather than starting with a pre-determined sampling frame. The interviewees were selected based on their particular knowledge and experience with empirical inquiry. It is also referred to as theoretical sampling. Snowball sampling also helped to a certain extent.

The advantage of this random sampling method is that it facilitates better causal inferences from the sample to the population, hence ensuring external validity of the findings. This is contingent on ensuring that the sample, although random, is still representative.

Sampling frame & sample matrix

The sampling frame of this research will constitute individuals with a relevant academic and professional background, who have had some form of interaction with algorithmic processing. Such interaction may have been either directly through the organisations they represent, or through their own study and work on the subject. Regarding the sample matrix, all interviewees will be of Indian nationality and based in India. While 83% of interviewees are male, 17% are female. All the interviewees are English speaking, as the interviews were conducted in English.

Thematic analysis

An integral method to analyse the data from the interviews is thematic analysis (TA). TA is a method to identify, interpret and analyse patterns of themes within qualitative data. This offers a toll (unbounded by theoretical commitments) rather than a methodology. Thus, it can be applied across a range of theoretical frameworks (Clarke & Braun, 2017).TA provides procedures to generate codes

¹² https://link.springer.com/referenceworkentry/10.1007%2F978-94-007-0753-5_2337#:~:text=Purposive%20sampling%20is%20intentional%20selection,theme%2C%20concept%2C%20or%20phenomenon.

and themes for qualitative data in a systematic and accessible manner. These codes are the smallest units of analysis which capture interesting features of the data that help answer the research question. They build themes which are the larger patterns of meaning and provide a framework for organizing and reporting observations (Clarke & Braun, 2017).

TAs are extremely useful as they allow flexibility in terms of the research question, sample, data collection methods and approaches to generate meaning. It can be used to identify meaning within and across data and can also be used within a theoretical framework (Clarke & Braun, 2017). Unlike conversation analysis, TA does not involve a micro analysis of language use and does not require a technical knowledge of language practice (Clarke & Braun, Thematic analysis, 2014).

A TA usually involves a six-phase process which is preceded by a critical reflection of the questions of the TA. The research question guides the coding and theme development, but the question can be modified and developed through the process. The first phase is when the researcher familiarises themselves with the data (through listening to audio recordings and reading interview transcripts). Notes are then taken and initial analytical observations for further exploration are specified. These codes are self-sufficient and the researcher does not need to refer to the raw data in later stages of the analysis. In the third stage, a set of themes is generated and the coded data is examined to identify overlaps. A figurative representation of the relationships between themes (thematic map of the analysis) is made. The themes are reviewed against the coded data in the fourth phase to ensure that there is a good fit between the two and that there is a coherent story about the coded data. Theme development is a recursive process until each theme is coherent and able to answer the research question meaningfully. Then, theme definitions are added to tell a story of each theme as well its central concept, scope, boundaries and relation to the other themes/ research question. The final phase is writing up the analysis which also involves assembling, editing, organising and further analysing the data (Clarke & Braun, Thematic analysis, 2014).

Potential interviewees

The interviewees will consist of a select group of 14 individuals in number. Participants will be chosen on the basis of their professional background and vantage point of engaging with algorithmic processing. While interviews with government and regulators could aid in understanding regulatory attitudes towards algorithmic processing, interviews with industry could shed light on how explainability is viewed within organisations. The composition of the interviewees will be as follows:

Group 1: Regulators (RG):

- a) Secretary or additional secretary level bureaucrats at the Ministry of Electronics and Information Technology (MeitY);
- b) Union minister of Electronics and Information Technology;
- c) Leadership level officials at NITI Aayog;
- d) Leadership level officials at regulators like RBI, SEBI, and IRDAI;
- e) Chairperson and/or members of the proposed Data Protection Authority of India to be set up under the PDP Bill, 2019;

Group 2: Private Sector (PS)

- f) Leadership position holders at prominent multinational and Indian technology companies;
- g) Leadership position holders at financial services entities deploying algorithms for automating decision-making;
- h) Individuals at the helm of MSMEs;
- i) Developers of algorithms.

Group 3: Users/ Researchers (UR)

- j) Influential users;
Representatives of user associations/groups

Analysis of interviews

Semi-structured interviews were conducted with fourteen participants, belonging to a wide spectrum of institutions cutting across private sector, think tanks, and academia. A thematic analysis of the interviews has been undertaken below, in order to identify relevant patterns to bolster our understanding of algorithmic explainability. A thematic analysis has the advantage of being flexible and explorative, and thus being suitable for analysis of qualitative data.

The range of interviews conducted yielded varying inputs that help provide a comprehensive and holistic overview of the status, role and future of AI in financial services in India. This overview is made up of several specific key insights, including into the permeation of AI services in Indian financial decision-making (opportunities and limitations observed), the evolution of the AI movement in the country (that it is still at an early, nascent stage), how regulatory authorities have kept up with this evolution (and the regulatory models available for adoption), and factors that set India apart from the rest (like the limited digital and financial literacy among citizens), among others.

In this regard, five themes have been identified for the analysis of these interviews. These are as follows: a) pervasion of ADS in the financial services sector, b) maturity of responsible AI movement in India c) appropriate mode of regulation, d) unique constraints operating in the Indian context, and e) trade-off between algorithmic explainability and algorithmic performance.

Pervasion of ADS in the financial services sector

The first theme relates to the extent of pervasion of ADS in the financial services sector, through services such as robo-advisory. This is helpful in understanding the areas within the financial services sector, or within fintech, which have seen the most uptick in adoption of AI/ML and ADS. Additionally, this theme sheds light on how far entities in both the public sector and the private sector are going to automate their processes in the financial services sector.

Most participants were of the unanimous view that there has been a noticeable upsurge in the pervasion of ADS in the financial services sector. One person (PS)

associated with the central bank of India said that the use of ADS is growing phenomenally. The reasons behind such an increase were discussed by the participants. A prominent reason was identified to be the massive volumes of fintech transactions being done, as well as increased spending capacity of the Indian middle class. In fact, one of the researchers (UR) was of the view that ADS will effectuate a dramatic shift in the lending sector in India — from a collateral-based model to a data-driven model. Therefore, banks must prepare suitably for this shift where algorithms would constitute the crux. In the opinion of one private sector representative, AI has different applications in various parts of the financial services value chain. The application of AI in the advisory part of the value chain has gained traction in the last 2-3 decades.

In their interview, one of the private sector (PS) representatives pointed out that automation in the investment space has been further propelled by the availability of rich data from social media engagement and customer transactions. This is several notches above the erstwhile data available from KYC questionnaires and customer demographic surveys. Another private sector (PS) representative pointed out a consumer preference that is emerging. As more people see the value of automated decision-making, they are gravitating towards AI-driven products. This is giving them more transparency and discipline in their financial decisions. Overall, the end-user trusts the ADS-driven philosophy of investing.

ADS infiltration in financial services: Factors at play



- On the lending side, the challenge there is money going out of the business. So organisations would prefer augmentative solutions rather than full replacement. The final decision-making will be left to a human
- India as a country is diverse and people expect the best services at the lowest prices. This will be a difficult balance to achieve
- There will be people who prefer the premium service of 'interaction with advisors' or private wealth managers



- Rapid adoption in FinTech because it is an area where the errors are generally perceived to be manageable, unlike healthcare
- Investment advisory has been dominated by cookie-cutter kind of products. These products have seen substantial adoption of ADS, which will increase further
- Volume of transactions as well as consumers' ability to accept conversational AI interactions for financial guidance will increase

Figure 1. 4 Positive (right column) and unfavourable (left column) factors at play in the infiltration of ADS in financial services

On the issue of where within the financial services sector the shift to ADS is most noticeable, participants had differing perspectives. This was perhaps due to varying levels of exposure to various parts of the financial services value chain. According to a private sector representative (PS), the use of ADS and robo-finance historically started out on the lending side, and was seen on the investment advisory side much later. On the other hand, a researcher (UR) was of the view that more automation is noticeable on the investment end, where algorithms can be taught to mimic experts and make decisions accordingly. The lending side, in contrast, is more challenging and organisations are not yet comfortable fully automating the process.

It was interesting to note the view of one of the researchers (UR) on fintech entities' attitude towards AI/ML. According to him, fintech is an area where the gains of adopting AI/ML are perceived to outweigh any potential harms. Such gains can be seen in the form of transactional efficiency and cost-cutting. In fact, another researcher (UR) supported this view, saying that ADS can bring in

transparency against social biases and this will be a big reason for higher adoption rates.

One researcher (UR) pointed out that while a lot of automation is being implemented, there is limited algorithmic intervention in investment. Another researcher (UR) pointed out that conflicts of interest will also play an important role in adoption of ADS in financial services. Old problems with respect to advisors being driven by the wrong incentives will continue to apply to AI-driven firms. This particular research seemed to carry a pessimistic view about the possibility to implementing robust regulations that help create a transparent and technologically advanced financial services sector in the future as perverse incentives of private players will continue to distort the market and come in the way of best practices development as more and more ADS is adopted.

Maturity of the responsible AI movement in India

The second theme seeks to put the finger on the interviewees' perception of the degree of maturity attained by the responsible AI movement in India until now. While the focus on responsible AI can be seen in academic literature, it is useful to take a step back and question how far such discourse has permeated on the ground. On this front, the responses from the interviewees prove to be reflective of the early stages in which we are operating, and provide a stark reminder of the work remaining to be done.

All the interviewees opined that the responsible AI movement in India remains at an early stage. The degree of maturity was described by various participants as "nascent", "a lot remaining to be done", and "very early stages". Nevertheless, some interviewees were quick to observe the potential of the responsible AI agenda. A number of participants referred to the NITI Aayog's paper on Responsible AI. A private sector representative observed that the draft titled "Responsible AI: #AIFORALL Approach Document for India" and released in February 2021, constitutes a significant development. As a contributor to the report, he was optimistic about various stakeholders coming together to achieve responsible AI in India.

One participant (PS) associated with the central bank noted that there are speculations relating to sectoral regulators setting up working group on responsible AI. According to one researcher (UR), even though the movement is at a fledgling state in India at present, there is scope for exponential explosion. They were also optimistic that responsible AI will feature as the core in discussions in private companies as well as regulators in the years to come. In fact, they proposed that algorithms themselves could serve as the solution for the problem of algorithmic bias. This may be done by fixing algorithms to eliminate any latent bias — a task that is easier for machine learning algorithms than human beings, as per him. Another private sector representative (PS), however, admitted that they had not yet had any discussions on the ethics of AI with any of their clients. They emphasised that in India concerns or issues around AI have not yet been heard of.

Another private sector representative (PS) mentioned that the conversation around responsible AI took off 6-9 months prior to the US elections. Questions were raised about whether automated decisions around insurance claims, for example, were biased. According to this participant, explainable AI will have to be a part of machine learning implementation. Regulatory questions or bias-related questions could then be answered and entities can become more responsible. The companies will need to take the onus themselves to offer explainability. This idea, however, is not mainstream yet.

On the other hand, an investment professional (PS) stated that scrutiny of algorithms from a responsible technology point of view is still some way away, and used the RBI committee on security of fintech apps as an illustrative example of how government is limited to ensuring basic security and not algorithmic accountability. The participant (PS) also predicted that civil society and media will continue to make noise about any malpractices. One researcher predicted that conversations around responsible AI will begin sometime next year when more AI-driven solutions will be deployed at scale.

One economist (PS) mentioned that in India, unethical behaviour is rampant in the financial services sector, and new technology will not change that unless

solutions are found and implemented through law and policy. Given that India will also have a lot of small value consumers on whom companies will continue to not spend too much money, the adoption of ADS in finance may throw up significant regulatory challenges.

Overall, while none of the respondents stated that responsible AI as a movement is catching up in quickly India as in the West, most pointed toward the *need* for this to happen. Given the vast data resources of India and the numerous ADS-driven applications being designed and implemented in the country, all respondents saw responsible AI gaining importance as a topic among relevant stakeholders in the coming months and years.

Appropriate mode of regulation

The third theme involves understanding the appropriate mode of regulation of AI/ML and ADS. Considering the nascence of the responsible AI agenda, and the fact that we are only now beginning to understand the potential harms of ADS, the issue of how best to regulate becomes critical. It can have significant implications for the manner in which AI/ML and ADS develops in future in India.

On this issue, participants were not willing to favour a purely self-regulatory model, as there seemed to be an impression that such a model would not be entirely effective in addressing the potential harms arising from ADS. One researcher (UR) opined that instead of self-regulation, a co-regulatory model formed on a harms-based approach would be a better fit.

Some participants believed that entities deploying AI/ML need guidance on how to incorporate AI governance. To this extent, a purely self-regulatory model would be inadequate in achieving algorithmic accountability. One private sector representative (PS) pointed out that regulators such as the RBI and SEBI should step in to regulate ADS in the financial services sector and bring in their expertise. Regulators (RG) should devote attention to adoption of ADS as it would affect the public at large and thus deserves regulatory attention.

With regard to the substantives of the regulation, another private sector representative (PS) offered an interesting suggestion. According to him, the

requirements imposed by regulators should be directly proportionate to the kind of information being processed by the entity. In other words, regulators should seek explainability only for product types with higher fiduciary obligations, and not across the board for all kinds of products.

One private sector representative (PS) mentioned that there may be some challenges in terms of the definition of 'robo-advisory' as well. Currently, mutual fund distributors and AI-powered financial advisors are clubbed together and there is need to re-think that categorization. Another private sector representative (PS) predicted that explainability will become a mandated compliance requirement. They think it will be in the natural progression of things that regulators will ask for more and more explainability. However, explainable AI will take time to become mainstream.

A representative of an industry body (RG) drew a parallel with the privacy law regime, wherein India's approach has been informed by the various Western approaches. In the AI field, the representative was hopeful that SEBI will take the lead as opposed to RBI, and conjectured that regulatory frameworks will not evolve in the immediate future. The representative (RG) was also averse to the idea of the proposed Data Protection Authority taking charge of the regulatory regime around AI in financial services. The investment professional (PS) mentioned that businesses would prefer audits of their algorithm as opposed to regulation. The person associated with the central bank (RG) also referred to audits as a way to regulate AI, and in that case, explainable AI will play an important role.

One researcher (UR) explicitly pointed out that state power to walk into private entities' offices and demand information on algorithms may lead to an undesirable situation. Another researcher (UR) pointed toward the toxic nature of internet businesses which are constantly focussed on more engagement and conjectured that perhaps only a non-profit cooperative-driven model can make entities reliable for the purpose of providing helpful financial advice, especially if they are deploying ADS.

REASONING FOR LIGHT-TOUCH REGULATION

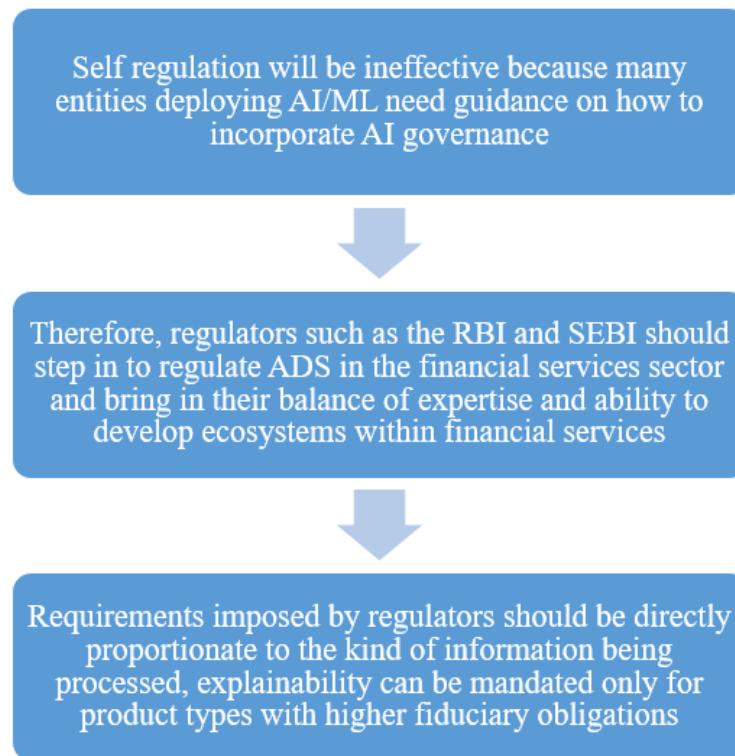


Figure 1. 5 Reasoning offered for light-touch regulation

Unique constraints operating in the Indian context

The present paper has earlier discussed the unique factors at play in India that assume relevance when regulating for algorithmic accountability. This is based on the understanding that while the global debate on the right to explainability can inform and shape India's policy, the ultimate formulation of the right must be determined by the country's unique circumstances. However, at least one of the private sector representatives pointed out that ethics of AI was not part of the discussion in their start-up, and their Indian consumers have not yet expressed any concerns or issues.

Most interviewees were cognisant of the nascence of the responsible AI agenda in India as of today. Moreover, some interviewees also pointed out that there does not exist a dedicated citizens' forum or group to air the voice of users. This is odd, considering the pervasion of AI/ML in applications that are used on a daily

basis by large swathes of the population. Citizens should thus be made more aware of the impact of generally applicable technology such as AI/ML and deep learning.

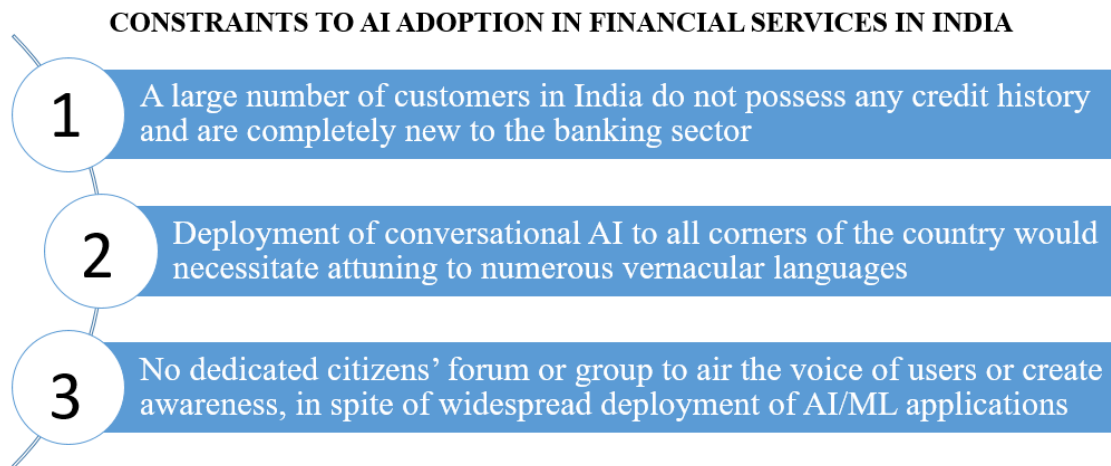


Figure 1. 6 Constraints cited for AI adoption in financial services in India

Making a significant point, a private sector representative (PS) highlighted that a large number of customers in India do not possess any credit history and are completely new to the banking sector. The presence of a sizeable unbanked segment will make it especially difficult to infuse algorithmic transparency. Further, another private sector representative (PS) pointed out that the presence of numerous local languages and dialects in India would present unique challenges to adoption of AI/ML and ADS. For example, deployment of conversation AI to all corners of the country would necessitate attuning to vernacular languages.

The person associated with the central bank (RG) mentioned that the problem in India is that subject matter expertise is inadequate both in terms of domain understanding of algorithms as well as finance. In the Indian context, interdisciplinary perspectives are not encouraged, and this will delay the process of effectively regulating AI.

One researcher (UR) pointed out that in India, mostly people invest in real estate and gold, besides shares and mutual funds. There is therefore not much going on in terms of ADS in financial services as the number of investors is very limited.

Trade-off between algorithmic explainability and algorithmic performance

The apparently inverse relationship between algorithmic explainability and algorithmic performance has been discussed earlier in the paper. It presents a critical juncture in the discourse surrounding algorithmic accountability and transparency. This is because it requires the clear prioritisation of one value over the other.

On this question, interviewees were sensitive to the gains and losses of choosing either of the values. Nevertheless, most interviewees considered explainability to be the most important goal – even at the cost of algorithmic performance. One researcher (UR) mentioned that to the extent that the adoption of explainability can avoid harms, it should be prioritised over performance and efficiency. One private sector representative (PS), however, pointed out that currently demand for explainability of AI-decisions was low, and they only receive such a request once in four months. When they do receive a request, they don't offer explainability because in case the consumer is financially qualified, then it would compromise their intellectual property, and in case the consumer is not financially qualified, then there is no point in revealing their data. They are completely reliant on the consumer's ability to trust them in making decisions. In terms of performance, their company offers more transparency.

A private sector representative (PS) explained how explainability is in the interests of all stakeholders – regulators, entities deploying ADS as well as individual users. While explainability would assist regulators in ensuring accountability from entities, it would also assist entities in avoiding hefty fines and complying with regulatory mandates. Users (UR) would also be benefited by the adoption of explainability. Another private sector representative (PS) was of the view that explainability should trump performance because algorithms would in any event perform better than human beings (even with diminished efficiency).

Another private sector representative (PS) differed from the rest in saying that the prioritisation should be left to the market forces, with consumers deciding what they desire more.

One private sector representative (PS) said that performance expectations will remain, but the idea will be to offer a degree of explainability as well. Explainability will not be expected to be offered on every data point that is being used. When bias is suspected, for example, not all data point will have to be explained. Therefore, unless that one data point is affecting performance, which is unlikely in case of neural networks, this will not be a trade-off.

On the other hand, one private sector representative (PS) asserted that many financial services players will not be willing to offer explainability because they may end up reveal too much of their intellectual property. In case the inability to offer explainability leads to loss in business, many entities will reconcile with that for competitive reasons. In fact, with respect to their company, the representative was clear that their products were only for people who are unable to invest by themselves and not for consumers interested in going into the details of specific decisions or making their decisions themselves. Their business will continue to depend on complete trust.

Another private sector representative (PS) state that if data is verified properly, explainability should not be required. Like negative testing is done for software lifecycle development, explainability will play the role of negative testing for machine learning applications, but a high level explainability will be difficult to offer.

One of the researchers (UR) pointed towards the fact that AI-driven businesses are funded by venture capital and consumers cannot be dependent on only companies or the industry to self-regulate effectively. In this context, the researcher (UR) mentioned that regulatory checks and punitive measures will be required eventually. In the absence of proper guidelines around parameters on which algorithms operate so that they can be explained, there can be no audits to understand why certain decisions are made.

The person associated with the central bank (RG) drew a parallel to green technologies, wherein the focus is not only on performance of technology, but also on sustainability. Similarly, the time is ripe for regulators to step in and ensure a culture of explainability is setup in the early days. Otherwise, there may be an irreversible trend in the direction of setting up high-performance algorithms that are not explainable. A researcher (UR) pointed out that as these solutions scale, the demands for explainability will increase. The researcher also drew attention to the fact that humans are also biased, and any explainability, or human intervention in the absence of explainability, has to be viewed relative to that.

One researcher (UR) drew a parallel to doctors who explain their treatment to patients and how they are not necessarily the best doctors. In the same way, AI explainability can be on the basis of black box testing, where the importance of relevant metrics can be gradually communicated to consumers.

Overall, many participants understood the inherent challenges in making certain kinds of AI-driven decision making to be made explainable. However, most participants cared about a certain level of transparency to be there around decisions and some participants were open-minded about this level of transparency being achieved at the cost of performance.

Table 1. 2 Comparing the observations on prioritising explainability versus prioritising performance

WEIGHING THE BENEFITS	
Prioritising explainability	Prioritising performance
<p>Today the comparison is between AI and humans. So, the focus should not be on doing things at the speed of light, but building trust. [Private sector representative]</p>	<p>In sectors like financial trading, performance has to be given preference over explainability. There is no penalty for an error beyond minor fines but organisations lose money if their speed (performance) is less. [Private sector representative]</p>

<p>Explainability will help create a more stable financial system, as desired by regulators. [Researcher and Investment Professional]</p>	<p>Theoretically speaking, explainability is seen in deontological terms. In the Indian context, we are looking at efficiency gains, and other parameters. Explainability should thus be seen as instrumental value. [Researcher]</p>
<p>Financial services organisations themselves will want more explainability primarily because the quantum of fines in financial services is really high and there is an emphasis on transparent decision making. [Private sector representative]</p>	<p>Users can always choose to work with human decision-makers if they want explainability. AI-driven products will add value because of superior performance and create efficiencies. [Private sector representative]</p>

Conclusion

The fintech sector has seen formidable growth and innovation in India. There has been ample support from policy-makers, with relevant institutions pushing for deepening of digital payments. Given the massive data resources of our country, it is likely that stakeholders will continue to strive for robust ecosystem development and the creation of new economic opportunities. For the most part, technology platforms are gradually moving toward the finance sector as they aggregate more users and start benefitting from network effects. In this context, it is important to acknowledge that automated decision-making systems will inevitably be a part of emerging ecosystems within fintech. As larger numbers of people participate in technology platforms, more innovation will take place and entities will find new ways to add value through machine learning and artificial intelligence.

As many of the interviewees have stated, it is not quite a question of *if* ADS will play a significant role in India's financial services sector. Rather, it is more a question of *when* this will happen. Most interviewees were of the opinion that there is reason to be cautiously optimistic about the fact that penetration of ADS in financial services will increase noticeably in the next few years. However, most interviewees believe that regulation will only catch up at a later stage once the penetration has reached a significant level. Some interviewees understand that there will be continued conversation in the media or among civil society on the need for regulatory intervention in order to make AI explainable. However, in terms of actual action on the ground, none of the interviewees see any significant development in the near term.

With respect to the appropriate mode of regulations, a number of interviewees mentioned that industry would be most comfortable with audits that render more transparency around the AI they are deploying. One interviewee vocalised a cautious view that drew from their understanding of existing malpractices within the financial services sector. According to them, similar malpractices may continue as newer technologies are introduced and in fact, issues around conflict of interest may remain unresolved for some time. This view is one that needs to

be weighed carefully by stakeholders as larger and newer sections of people enter the financial services market and use ADS-driven products.

A number of interesting perspectives have emerged as a response to questions around the trade-off between explainability and performance. Industry representatives stated that private sector entities will optimise for explainability in many instances before prioritising performance. Participants seemed to be inclined toward explainability being a priority because of their awareness about the quantum of progress that can be made in terms of performance as ADS-driven products become a reality. One view was that this trade-off will play out differently in different parts of the value chain within the financial services sector.

Overall, the interviews revealed a cautiously optimistic view among stakeholders in terms of increased ADS penetration, its proper regulation and the increase in prominence of explainable AI. With a concerted effort from policy-makers to ensure the success of the 'Digital India' mission, it is likely that the private sector will step up on efforts to innovate and bring more users into the market.

Therefore, as the market develops, the picture with respect to consumer welfare and harm reduction from ADS-driven products will become clearer and regulators will step in accordingly. Similar to the case with privacy laws and other regulations around technology, Indian policy-makers will likely consider some approaches from the West first, before launching consultations internally and coming up with an appropriate regulatory solution that takes into account the unique challenges of this market.

Appendix (A)

List of Stakeholders

Group	Abbreviation	# of Interviews Conducted
Private Sector	PS	6
Regulators	RG	2
Users/ Researchers	UR	6

Frameworks to assess algorithms

FAT/ML; FATES

The Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) framework was established in 2014 in response to the growing recognition of the challenges arising from machine learning. It has functioned as an annual event conducted since its inception, where researchers explore how to address these issues through “computationally rigorous methods” (FAT/ML, 2014).

The five key principles of the FAT/ML framework are responsibility, explainability, accuracy, auditability, and fairness. Additionally, the framework proposes that algorithm creators develop a social impact statement hinged on the above principles. This would ensure adherence to the principles and display a public commitment to associated best practices. The social impact statement would contain answers to questions indicated by the framework, such as “[w]ho is responsible if users are harmed by this product?” and “[h]ow much of your system / algorithm can you explain to your users and stakeholders?”. The framework also delineates initial steps to be taken under each of the key principles, thus aiding algorithm creators in implementing the principles.

The research produced by the organisation relates to discrimination-aware data mining, discrimination-free classification, and responsible innovation of data mining and profiling tools. The scholarly work produced at these conferences, and the technical workshops held by FAT/ML, has been significant in starting and advancing the debate on ethical AI. Further movements have expanded the ambit

of the 'FAT' framework. For instance, the Microsoft research FATE group added the 'E' for ethics to FAT.

Later the 'FATE' framework has also been expanded to 'FATES', with an 'S' added to denote safety and security, in an attempt to capture all relevant parameters within the acronym (Wing, 2018). Safety and security refer to the need to ensure that engineered systems should do no harm and protect against malicious behaviour (Wing, 2018). If systems are not safe and secure, attaining consumer trust would be an uphill task. It is also crucial that the FATES framework is incorporated into a system prior to deployment, instead of post deployment (Wing, 2018).

Fairness, Accountability, Transparency and Ethics (FATE)

Fairness, Accountability, Transparency and Ethics (FATE) as a succinct framework for studying algorithmic decision making has become increasingly popular over the last few years. The framework received attention from industry and media when Microsoft, which is heavily reliant on AI/ML for its products, launched a research project by its name in 2014.

The framework may be deconstructed as follows. Fairness refers to models making unbiased decisions, accountability refers to assigning responsibility for a machine-made decision, transparency refers to being open to the end user about how a decision has been made, and ethics refers to paying attention to ethical and privacy-preserving uses of data, and the ethical decisions that require consideration in building an automated system. The project was focussed on promoting computational techniques that are simultaneously ethical and innovative, while at the same time being informed by the surrounding social and historical context (Microsoft).

The FATE project at Microsoft sought to understand the following questions:

- "How can AI assist users and offer enhanced insights, while avoiding exposing them to discrimination in health, housing, law enforcement, and employment?"

- How can we balance the need for efficiency and exploration with fairness and sensitivity to users?
- As our world moves toward relying on intelligent agents, how can we create a system that individuals and communities can trust?”

In this regard, the project has conducted research on the role human beings can play for intelligible machine learning, mitigation of bias in hiring decisions, and operationalising data minimisation in the context of personalisation. In order to achieve its goals under FATE, Microsoft has worked with research institutions like the AI Now Society at New York University and Partnership on AI. Since the launch of the Microsoft project, the FATE framework has spread in its reach due to its innate promise of putting forth a set of clear parameters to assess algorithms.

The appeal of the FATE framework lies in its crystallisation of the debate till date on algorithmic decision-making systems. It provides a guiding light to developers of algorithms and entities deploying algorithms on the broad criteria they should be attentive to while creating and auditing algorithms. However, dedicated attention is required to lend sufficient clarity to the manner in which the FATE framework is applied, and the precise contours of such application. The absence of such clarity may exacerbate the confusion arising from the regulatory vacuum on issues of algorithmic governance in most jurisdictions across the world.

Further, it is worth considering the manner in which incorporating ‘explainability’ within the ambit of FATE would be helpful in achieving the constituent elements of fairness, accountability, transparency and ethics. Not only would explainability fit in neatly with the FATE goals, it would also imbue an additional degree of conceptual cohesion in an otherwise relatively nebulous concept. More importantly, the inclusion of explainability within the framework would provide a practical means of achieving the avowed objects. Thus, while the FATE framework at present is useful in providing an overall sense of direction to algorithmic assessments, it is the brass tacks and future additions that will cement its role as a legitimate enabler of individual privacy and trust.

ETHICA

ETHICA is a risk-alleviation and de-biasing framework to achieve transparency and trust in automation floated by the Indian multinational corporation Wipro in 2019. It stands for Explainability, Transparency, Human-first, Interpretability, Common sense and Auditability. Wipro's State of Automation Report 2019 claims that ETHICA renders automation intelligible by taking it out of its black box (Wipro, State of Automation 2019, 2019). ETHICA is used by Wipro to power its automation and cognitive services platform Wipro HOLMES (Wipro, State of Automation 2019, 2019).

An effort to address concerns raised in the ethical AI debate, the six components of the ETHICA framework are intended to de-bias AI models from the perspective of operation and function. In this regard, some of the approaches it uses are masking of data, deploying ethics transparency and explainability as part of the development process, using proper anomaly detection, and human-based auditing (Wipro, State of Automation 2019, 2019). Reports indicate that Wipro intends on making the framework available to clients as a part of its bundle of services (Bureau, 2020). The goal of the framework is to bake in values of integrity, explainability and fairness within the design of automated systems, so as to boost consumer trust.

However, Wipro has admitted that de-biasing AI models and learning algorithms is no mean task. Cognitive automation is riddled with challenges involving both people and technology. For instance, deep learning algorithms require such large volumes of data that it is difficult to explain why an intelligent system arrived at a particular outcome. This has given impetus to the field of Explainable AI ('XAI'), which is intended to cull out rationales of automated decisions by using a FAT/ML model.

Further, the endeavour to alleviate AI risks may run into roadblocks as manually tuning algorithms is a time-consuming process and projects may await action for prolonged durations. It is also unknown how a particular algorithm would behave in a real-world scenario when it has to interact with other predictive algorithms (Wipro, The need for AI to sense, think, respond and learn without bias).

The ETHICA framework illustrates that key players in the information and communications technology sector may find it wise to develop their own in-house solutions for algorithmic governance. Development of ethical AI frameworks could establish responsible entities as leaders in the space. Moreover, such an approach enjoys the advantage of being highly customisable, relatively flexible vis-à-vis off-the-shelf solutions and more relevant to the organisation's goals and culture.

Wipro's move may encourage other software players to also develop their own ethical AI frameworks, and lead to a trend of industry leaders paving the path for ethical AI. It would be interesting to observe how such in-house frameworks serve to achieve ethical AI in the absence of overarching regulation on algorithmic governance. It however remains to be seen how these industry frameworks would fare in the face of government regulation, as and when it comes in.

Artificial Intelligence for Social Good ('AI4SG')

Another framework that has attained popularity is Artificial Intelligence for Social Good (AI4SG). The idea is aimed at advancing artificial intelligence to address societal issues and promote well-being (Shi, 2020). It is based on the realisation of social impact in accordance with the goals drawn out by the United Nations' 17 Sustainable Development Goals or SDGs (Tomasev, 2020). The movement is geared towards the forging of interdisciplinary partnerships in order to apply AI towards SDGs (Tomasev, 2020). Efforts at channelling AI4SG have sprung up in fields like climate informatics, monitoring of viral diseases, and prediction of poverty.

However, the primary flaw of the framework lies in the lack of clarity about the precise components of 'social good' – a concept that straddles ethics, law and social science. Neither 'AI' nor 'social good' have universally accepted definitions. This has encouraged an ad-hoc approach, where the application of AI in specific areas, like disaster management and health, is explored without explaining how AI4SG solutions should be designed to leverage the potential of AI (Floridi, 2020). This is especially critical considering many of the projects under this umbrella may have significant ethical implications.

Conceptual vagueness about the foundation of the AI₄SG framework may lead to iatrogenic scenarios, where an AI initially intended for social good ends up causing unintended harms (Floridi, 2020). Thus, context-specific design and deployment is required to ensure the successful delivery of AI₄SG projects (Floridi, 2020).

In this vein, Floridi has suggested seven essential factors for successful AI₄SG. These factors are as follows: a) falsifiability and incremental deployment; b) safeguards against the manipulation of predictors; c) receiver-contextualised intervention; d) receiver-contextualised explanation and transparent purposes; e) privacy protection and data subject consent; f) situational fairness; and g) human-friendly semanticisation (Floridi, 2020). The above factors are posited to be applied in conjunction with the five principles of ethical AI, namely beneficence, nonmaleficence, autonomy, justice and explicability, in order to achieve social good through AI (Floridi, 2020). Furthermore, guidelines have been proposed for successful AI₄SG collaborations. These include ensuring that AI applications are inclusive and accessible, reviewed at every stage for ethics and human rights compliance, and defining goals and use-cases in a clear manner (Tomasev, 2020). The suggestions furnished above may lend conceptual coherence to the AI₄SG framework, thus enabling better realisation of its goals.

References

- Aayog, N. (2018). *National Strategy for Artificial Intelligence*. New Delhi.
- Act, A. A. (2019). Retrieved from
<https://www.wyden.senate.gov/imo/media/doc/Algorithmic%20Accountability%20Act%20of%202019%20Bill%20Text.pdf>
- Act, D. P. (2018).
- Andrew D Selbst, J. P. (2017). *Meaningful information and the right to explanation*. Retrieved from International Data Privacy Law :
<https://academic.oup.com/idpl/article/7/4/233/4762325>
- Andrews, L. (2017). *Algorithmic Regulation*. London: LSE.
- APRU. (2020). *Artificial Intelligence for Social Good*. Retrieved from
https://apru.org/wp-content/uploads/2020/09/layout_v3_web_page.pdf
- (1950). Article 14. In *Constitution of India*.
- Baldwin, R., Cave, M., & Lodge, M. (2012). *Understanding Regulation: Theory, Strategy, and Practice*. Oxford: OUP.
- Bank, W. (2019, February). *Robo-Advisors: Investing through Machines*. Retrieved from
<http://documents1.worldbank.org/curated/en/275041551196836758/pdf/Robo-Advisors-Investing-through-Machines.pdf>
- Basu, A. (2018). Discrimination in the Age of Artificial Intelligence. *Oxford Human Rights Hub*.
- Basu, A. (2019). *What is the problem with 'Ethical AI'? An Indian Perspective*. Retrieved from CIS: <https://cis-india.org/internet-governance/blog/what-is-the-problem-with-2018ethical-ai2019-an-indian-perspective>
- Bryman, A. (1984). The Debate about Quantitative and Qualitative Research: A Question of Method or Epistemology? *The British Journal of Sociology*.
- Bureau, E. T. (2020, July 28). *Wipro may offer AI tool Ethica to clients as the CEO looks to revamp digital technology-based solutions*. Retrieved from

Economic Times:

<https://economictimes.indiatimes.com/tech/software/wipro-may-offer-ai-tool-ethica-to-cos/articleshow/77206871.cms?from=mdr>

Canada, G. o. (2019). *Directive on Automated Decision-Making, Appendix A.*

Retrieved from <https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592>

Canada, G. o. (2019). *Minister Bains announces Canada's Digital Charter.*

Retrieved from <https://www.canada.ca/en/innovation-science-economic-development/news/2019/05/minister-bains-announces-canadas-digital-charter.html>

Canada, G. o. (2020). *Algorithmic Impact Assessment.* Retrieved from

<https://canada-ca.github.io/aia-eia-js/>

Canada, O. o. (2019). *PIPEDA fair informational principles.* Retrieved from

https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/p_principle/

Canada, O. o. (2019). *PIPEDA in brief.* Retrieved from

https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/pipeda_brief/

Canada, O. o. (2020). *Consultation on the OPC's Proposals for ensuring appropriate regulation of artificial intelligence.* Retrieved from

https://www.priv.gc.ca/en/about-the-opc/what-we-do/consultations/consultation-ai/pos_ai_202001/

Centre for Data Ethics and Innovation. (2019). *Interim report: Review into bias in algorithmic decision-making.* London: UK Government.

Chatterjee, S., & Ravindran, S. (2019). *There is need for a legal, organisational framework to regulate bias in algorithms.* Retrieved from The Indian

Express: <https://indianexpress.com/article/opinion/columns/rules-for-the-machine-algorithm-artificial-intelligence-5603795/>

- City of Somerville, M. (2019). *Banning the usage of facial recognition technology in Somerville*. Retrieved from http://somervillecityma.iqm2.com/Citizens/Detail_LegiFile.aspx?Frame=&MeetingID=2926&MediaPosition=2219.261&ID=20991&CssClass=
- Commission, A. C. (2017). *Consumer Data Right*. Retrieved from <https://www.accc.gov.au/focus-areas/consumer-data-right-cdr-o>
- Commission, A. H. (2019). *Human Rights and Technology: Discussion Paper*. Retrieved from https://tech.humanrights.gov.au/sites/default/files/2019-12/TechRights2019_DiscussionPaper.pdf
- Commission, A. H. (n.d.). *Human Rights and Technology, Our Work*. Retrieved from <https://tech.humanrights.gov.au/our-work>
- Commission, E. (2019). *Ethics guidelines for trustworthy AI*. Retrieved from <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>
- Commission, E. (2020). *White Paper: On artificial intelligence- A European approach to excellence and trust*. Retrieved from https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
- Council, T. N. (2017). *A Local Law in relation to automated decision systems used by agencies*. Retrieved from <https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2-9C42-461253F9C6Do&Options=&Search>
- Crawford, K., & Schutz, J. (2019). AI systems as state actors. *Columbia Law Review*, 119(7).
- CSIRO. (2019). *Artificial Intelligence Australia's Ethics Framework: A Discussion Paper*. Retrieved from https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdf

- Drozdiak, N. (2020). *Even the Pandemic Doesn't Stop Europe's Push to Regulate AI*. Retrieved from <https://www.bloomberg.com/news/articles/2020-04-07/coronavirus-isn-t-stopping-europe-s-push-to-regulate-ai>
- Edwards, L., & Veale, M. (2017). Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For. *Duke Law & Technology Review*.
- Engler, A. (2020). *The European Commission considers new regulations and enforcement for "high-risk" AI*. Retrieved from Brookings: <https://www.brookings.edu/blog/techtank/2020/02/26/the-european-commission-considers-new-regulations-and-enforcement-for-high-risk-ai/>
- FAT/ML. (2014). *Fairness, Accountability, and Transparency in Machine Learning*. Retrieved from <https://www.fatml.org/>
- Fjeld, J. (2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*. Berkman Klein Center for Internet and Society at Harvard University.
- Fjeld, J., & Nagy, A. (2020, January 15). *Principled Artificial Intelligence*. Retrieved from Harvard Berkman Klein Center for Internet and Society at Harvard University: <https://cyber.harvard.edu/publication/2020/principled-ai>
- Floridi, L. (2020). *How to Design AI for Social Good: Seven Essential Factors*. Retrieved from <https://link.springer.com/article/10.1007/s11948-020-00213-5>
- Foundation, E. F. (2019). *Oakland Face Surveillance Ban*. Retrieved from <https://www.eff.org/document/oakland-face-surveillance-ban>
- Francisco, S. (2019). *Stop Secret Surveillance Ordinance*. Retrieved from <https://sfgov.legistar.com/View.ashx?M=F&ID=7206781&GUID=38D37061-4D87-4A94-9AB3-CB113656159A>

- FTC. (2016, January). *Big Data: A Tool for Inclusion or Exclusion?* Retrieved from <https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf>
- FTC. (2018, November). *Competition and Consumer Protection Implications of Algorithms, Artificial Intelligence, and Predictive Analytics*. Retrieved from https://www.ftc.gov/system/files/documents/public_statements/1431041/hoffman_-_ai_intro_speech_11-14-18.pdf
- FTC. (2018). *Privacy and Data Security Update: 2018*. Retrieved from <https://www.ftc.gov/system/files/documents/reports/privacy-data-security-update-2018/2018-privacy-data-security-report-508.pdf>
- FTC. (2020, April). *Using Artificial Intelligence and Algorithms*. Retrieved from <https://www.ftc.gov/news-events/blogs/business-blog/2020/04/using-artificial-intelligence-algorithms>
- GDPR, E. (2018).
- Gillis, T., & Spiess, J. (2019). Big Data and Discrimination. *The University of Chicago Law Review*, 86(2).
- Government, U. (2019). *Press release: Investigation launched into potential for bias in algorithmic decision-making in society*. Retrieved from <https://www.gov.uk/government/news/investigation-launched-into-potential-for-bias-in-algorithmic-decision-making-in-society>
- Heaven, W. D. (2020, January). *Why asking an AI to explain itself can make things worse*. Retrieved from MIT Technology Review: <https://www.technologyreview.com/2020/01/29/304857/why-asking-an-ai-to-explain-itself-can-make-things-worse/>
- Hickok, E., & Basu, A. (2018). *Artificial Intelligence in the Governance Sector in India (Working Draft)*. Retrieved from Centre for Internet and Society: <https://cis-india.org/internet-governance/ai-and-governance-case-study-pdf>

- Hunt, R., & McKelvey, F. (2019). Algorithmic Regulation in Media and Cultural Policy: A Framework to Evaluate Barriers to Accountability. *Journal of Information Policy*, 9, 307-335.
- India, I. a. (2019). *India Internet 2019*. Retrieved from <https://cms.iamai.in/Content/ResearchPapers/d3654bcc-002f-4fc7-ab39-e1fbeb00005d.pdf>
- India, T. G. (2016). *THE AADHAAR (TARGETED DELIVERY OF FINANCIAL AND OTHER SUBSIDIES, BENEFITS AND SERVICES) ACT, 2016*. Retrieved from https://uidai.gov.in/images/targeted_delivery_of_financial_and_other_subsidies_benefits_and_services_13072016.pdf
- India, U. I. (n.d.). *Aadhaar*. Retrieved from <https://uidai.gov.in/>
- Information Commissioner's Office, U. (2017). *Big data, artificial intelligence, machine learning and data protection*. Retrieved from <https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf>
- Information Commissioner's Office, U. (2019). *Human Bias and Discrimination in AI systems*. Retrieved from <https://ico.org.uk/about-the-ico/news-and-events/ai-blog-human-bias-and-discrimination-in-ai-systems/>
- Information Commissioner's Office, U., & Institute, A. T. (2019). *Explaining decisions made with AI*. Retrieved from <https://ico.org.uk/media/about-the-ico/consultations/2616434/explaining-ai-decisions-part-1.pdf>
- Institute, T. A. (2018). *A Right to Explanation*. Retrieved from <https://www.turing.ac.uk/research/impact-stories/a-right-to-explanation>
- IRDAI. (2017). *(Protection of Policyholders' Interests) Regulations, Regulation 19(5)*.
- IRDAI. (2017). *Cyber Security Guidelines*.
- IRDAI. (2017). *Guidelines on insurance e-commerce*.

- Joshi, D. (2020). *India's privacy law needs to incorporate rights against the machine*. Retrieved from Medianama:
<https://www.medianama.com/2020/05/223-indias-privacy-law-needs-to-incorporate-rights-against-the-machine/>
- Joshi, D. (2020). Welfare Automation in the Shadow of the Indian Constitution. *Socio-Legal Review*.
- K.S. Puttaswamy v Union of India, (2019) 1 SCC 1 (2019).
- K.S. Puttaswamy v. Union of India (majority opinion delivered by Chandrachud J.), 2017 10 SCC 1 (SC 2017).
- Kearns, M., & Roth, A. (2020, January 13). *Ethical algorithm design should guide technology regulation*. Retrieved from Brookings:
<https://www.brookings.edu/research/ethical-algorithm-design-should-guide-technology-regulation/>
- Knight, W. (2017, April 11). *The Dark Secret at the Heart of AI*. Retrieved from MIT Technology Review:
<https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/>
- Krishnan, S., Deo, S., & Sontakke, N. (2020). *Operationalizing algorithmic explainability in the context of risk profiling done by robo financial advisory apps*. Data Governance Network.
- Kumar, R. (2019). *India needs to bring an algorithm transparency bill to combat bias*. Retrieved from ORF: <https://www.orfonline.org/expert-speak/india-needs-to-bring-an-algorithm-transparency-bill-to-combat-bias-55253/>
- Lee, N. T., Resnick, P., & Barton, G. (2019, May 22). *Algorithmic bias detection and mitigation: best practices and policies to reduce consumer harms*. Retrieved from Brookings:
<https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/>

- Lessig, L. (n.d.). *Code is law*. Retrieved from Harvard Magazine:
<https://harvardmagazine.com/2000/01/code-is-law-html>
- Lofchie, S. (2020, February 19). *United States: House Committee Considers Measures To Reduce AI Bias In Financial Services*. Retrieved from Mondaq:
<https://www.mondaq.com/unitedstates/financial-services/894982/house-committee-considers-measures-to-reduce-ai-bias-in-financial-services>
- Lords, H. o. (2017). *AI in the UK: ready, willing and able?* Retrieved from
<https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm>
- Malgieri, G. (2019). Automated decision-making in the EU Member States: The right to explanation and other “suitable safeguards” in the national legislations,. *Computer Law and Security Review*, 35(5).
- Marda, V. (2018). Artificial Intelligence Policy in India: A Framework for Engaging the Limits of Data-Driven Decision-Making. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*.
- Massachusetts, C. o. (2019). *An Act establishing a moratorium on face recognition and other remote biometric surveillance systems*. Retrieved from
<https://malegislature.gov/Bills/191/SD671>
- Merwe, D. M. (2010). *A consumer perspective on food labelling: Ethical or not?* Retrieved from Koers - Bulletin for Christian Scholarship:
https://www.researchgate.net/publication/274658213_A_consumer_perspective_on_food_labelling_Ethical_or_not
- Microsoft. (n.d.). *FATE: Fairness, Accountability, Transparency, and Ethics in AI*. Retrieved from <https://www.microsoft.com/en-us/research/theme/fate/>
- Mittelstadt, B. (2019). *Principles Alone Cannot Guarantee Ethical AI*. Retrieved from SSRN:
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3391293

- Mosley, L. (2013). *Interview Research in Political Science*. Cornell University Press.
- Ontario, I. a. (2017). *Big Data Guidelines*. Retrieved from <https://www.ipc.on.ca/wp-content/uploads/2017/05/bigdata-guidelines.pdf>.
- Party, A. 2. (2018). *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, 2017, as last revised on 6 February 2018*. Retrieved from https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053
- Pasquale, F. (2016). *Black Box Society*. Harvard University Press.
- (n.d.). Personal Data Protection Bill, 2019.
- Privacy Act, A. (1988).
- PwC. (2019, October 14). *PwC Global Fintech Report 2019: UK financial services firms trailblazing on automation efforts*. Retrieved from PwC: <https://www.pwc.co.uk/press-room/press-releases/uk-fintech-pwc.html>
- RBI. (2002). *Annexure: Information Systems Security Guidelines for the Banking and Financial Sector (Part 1 of 2)*. Retrieved from <https://www.rbi.org.in/Scripts/PublicationReportDetails.aspx?ID=277>
- RBI. (2017). *Report of the Household Finance Committee*. Retrieved from <https://www.rbi.org.in/Scripts/PublicationReportDetails.aspx?UrlPage=&ID=877#AF>
- RBI. (2018). *Storage of Payment System Data*. Retrieved from <https://www.rbi.org.in/Scripts/NotificationUser.aspx?Id=11244&Mode=0>
- SEBI. (2019). *Guidelines for Seeking Data*. Retrieved from <https://www.sebi.gov.in/pdf/guidelines-of-data-sharing-final.pdf>

- SEBI. (2013). *Memo on SEBI (Investment Advisers) Regulations, 2013*. Retrieved from https://www.sebi.gov.in/sebi_data/meetingfiles/mar-2020/1583318232255_1.pdf
- SEBI. (n.d.). SEBI (Investment Advisers) Regulations, 2013. In *Last amended on April 17, 2020*.
- Sharma, A. (2020). 'Uncertainty looming over Covid crisis bottoming out for MSMEs to assess need for more govt stimulus'. Retrieved from Financial Express: <https://www.financialexpress.com/industry/sme/cafesme/msme-eodb-uncertainty-looming-over-covid-crisis-bottoming-out-for-msmes-to-assess-need-for-more-govt-stimulus/1995517/>
- Shi, Z. R. (2020). Retrieved from <https://arxiv.org/pdf/2001.01818.pdf>
- Solove, D. J. (2013). *INTRODUCTION: PRIVACY SELF-MANAGEMENT AND THE CONSENT DILEMMA*. Retrieved from Harvard Law Review: https://harvardlawreview.org/wp-content/uploads/pdfs/vol126_solove.pdf
- Srikrishna, C. o. (2017). *White Paper of the Committee of Experts on a Data Protection Framework for India*. New Delhi: Ministry of Electronics and Information Technology.
- Srikrishna, C. o. (2018). *A Free and Fair Digital Economy: Protecting Privacy, Empowering Indians*. New Delhi: Ministry of Electronics and Information Technology, India.
- Srivastava, S. (2020). *India Must Treat the Internet as a Public Utility During COVID 19, and After*. Retrieved from The Wire: <https://thewire.in/tech/india-must-treat-the-internet-as-a-public-utility-during-covid-19-and-after>
- States, O. o. (2016). *The National Artificial Intelligence Research and Development Strategic Plan*. Retrieved from https://www.nitr.gov/pubs/national_ai_rd_strategic_plan.pdf

- Tomasev, N. (2020). *AI for social good: unlocking the opportunity for positive impact*. Retrieved from <https://www.nature.com/articles/s41467-020-15871-z>
- Turek, D. M. (n.d.). *Explainable Artificial Intelligence (XAI)*. Retrieved from DARPA: <https://www.darpa.mil/program/explainable-artificial-intelligence>
- Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2).
- Washington, S. o. (2019). *House Bill 1655*. Retrieved from <http://lawfilesexternal.leg.wa.gov/biennium/2019-20/Pdf/Bills/House%20Bills/1655.pdf>
- Wing, J. M. (2018). *Data for Good: FATES, Elaborated*. Retrieved from <https://www.datascience.columbia.edu/FATES-Elaborated>
- Wipro. (2019). *State of Automation 2019*.
- Wipro. (n.d.). *The need for AI to sense, think, respond and learn without bias*. Wipro.
- Zarsky, T. (2015). The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making. *Science, Technology and Human Values*, 41(1), 118-132.
- Ziewitz, M. (2016). Governing Algorithms: Myth, Mess and Methods. *Science, Technology and Human Values*, 41(1).

Operationalising algorithmic explainability in the context of risk profiling done by robo financial advisory apps

Abstract

Robo advisors are financial advisory apps that profile users into risk classes before providing financial advice. This risk profiling of users is of functional importance and is legally mandatory. Irregularities at this primary step will lead to incorrect recommendations for the users. Further, lack of transparency and explanations for these automated decisions makes it tougher for users and regulators to understand the rationale behind the advice given by these apps, leading to a trust deficit. Regulators monitor this profiling but possess no independent toolkit to “demystify” the black box or adequately explain the decision-making process of the robo financial advisor.

Our paper proposes an approach towards developing a ‘RegTech tool’ that can explain the robo advisor’s decision making, this explanation methodology can be extended to other complex algorithmic decision-making systems. We use machine learning models to recognise and reconstruct three levels of explanations, revealing the original risk profiling decision logic of the robo advisor. First, we find the importance of inputs used in the risk profiling algorithm. Second, we infer relationships between inputs and with the assigned risk classes. Third, we allow regulators to explain decisions for any given user profile, in order to ‘spot check’ a random data point. With these three explanation methods, we provide regulators, who lack the technical knowledge to understand algorithmic decisions, a method to understand it and ensure that the risk-profiling done by robo advisory applications comply with the regulations they are subjected to.

Keywords: Algorithmic decision-making systems (ADS), algorithmic regulation, algorithmic explainability and transparency, robo financial advisory apps, fintech, explainable AI

Introduction

There is a growing ubiquity of decision-making algorithms that affect our lives and the choices we make. These algorithms curate our internet and social media feed, trade in the stock market, assess risk in banking, fintech and insurance, diagnose health ailments, predict crime prevention, and a lot more. Broadly, these are known as Algorithmic Decision-making Systems (ADS). Machine learning algorithms are one of the crucial components of ADS and artificial intelligence (AI), and power the automated, independent decision making done by computers. Machines 'learn' by going through millions of data points and find associations and patterns in them. They then apply the learnt rules on new data to predict the outcomes. These algorithms have promised and delivered considerable gains in efficiency, economic growth, and have transformed the way we consume goods, services, and information.

However, along with the gains, these algorithms also pose threats. Several cases have come to light where algorithm powered decisions have given rise to undesirable consequences. An automated hiring tool used by Amazon discriminated heavily against women applying for software development jobs, because the machines learn from past data that has a disproportionate number of men in software positions (Dastin, 2018). Software used for crime prediction in the United States showed a machine bias against African-Americans, exacerbating the systemic bias in the racial composition of prisons (ProPublica, 2016). Google's online advertising system displayed ads for high-income jobs to men much more often than it did to women (Datta, Tschantz, & Datta, 2015). Social media algorithms are found to inadvertently promote extremist ideology (Costello, Hawdon, Ratliff, & Grantham, 2016) and affecting election results (Baer, 2019). Recently, researchers found that racial bias in the US health algorithms reduced the number of Black patients identified for extra care by more than half (Obermeyer, Powers, Vogeli, & Mullainathan, 2019) (Kari, 2019).

In effect, contrary to the promise of unbiased and objective decision making, these examples point to a tendency of algorithms to unintentionally learn and reinforce undesired and non-obvious biases, thus creating a trust deficit. This

arises mainly because several of these algorithms are not adequately tested for bias and are not subjected to external due-diligence. The complexity and opacity in the algorithms decision-making process and the esoteric nature of programming denies those affected by it access to explore the rights-based concerns posed by algorithms.

However, if these algorithms make decisions in the public sphere that affect an individual's access to services and opportunities, they need to be scrutinised. Over the last two years, there has been a growing call to assess algorithms for concepts like fairness, accountability, transparency, and explainability and there has been an increase in research efforts in this direction. These are all concepts that also came up in the first paper in this series, which studied the trade-off that emerges between algorithmic performance and explainability. The research in this second paper is situated in the same context, where we attempt to operationalise the concept of **explainability** in automated tools used in fintech. We have selected the case of **robo financial advisory apps** which conduct a **risk profiling** of users based on a questionnaire and gives users customised investment advice.

What are robo financial advisors?

Robo advisory applications are automated web-based investment advisory algorithms that estimate the best plans for trading, investment, portfolio rebalancing, or tax saving, for each individual as per their requirements and preferences. Typically, a user fills in questionnaire or survey and is classified in either three or five risk classes (ranging from 'low risk' to 'high risk'). Robo advisors open up the potential for finance to be democratised by reducing the financial barrier to entry and providing equal access to financial advice through their low-cost business model (Laboure & Braunstein, 2017).

The first robo financial advisory app was launched in 2008, and the use of such applications has increased with the growth of internet-based technology and the sophistication of functionalities and analytics (Abraham, Schumkler, & Tessada, 2019) (Narayanan, 2016). In a 2014 report, the International Organization of Securities Commissions (IOSCO) made a comprehensive effort to understand

how investment intermediaries use automated advisory tools. They identified a spectrum of 'Internet-based automated investment selection tools' and classified them based on the complexity of the advice that it gives, from a basic level of risk classification to a complex assessment of the customer's age, financial condition, risk tolerance, and capacity, among others, to offer automated advice suited to the user's investment goals. The output is often a set of recommendations for allocations based on parameters like the size of funds (small, mid-cap), the type of investment (debt and equity funds), and even a list of securities or portfolios (IOSCO, 2014).

This **risk profiling** done by these robo financial advisors is a crucial step to determine the risk class of the user which determines the investment advice. Irregularities at this primary step will lead to incorrect recommendations for the users. Moreover, unlike human advisors, robo advisors provide no reasons or explanations for their decisions, and this shortcoming reduces the trust that users repose in their advice (Maurell, 2019).

Several robo financial advisory applications operate in India. Prominent ones include PayTM money, GoalWise, Artha-Yantra, Upwardly, Kuvera, Scripbox, MobiKwick, and Tavaga, among others.

Regulating ADS

(Citron & Pasquale, 2014) argue that transparency and opening the black box are crucial first steps and that oversight over algorithms should be a critical aim of the legal system. They argue for procedural regularity in assessing all publicly used algorithms to ensure fairness.

The European Union General Data Protection Regulation (EU GDPR) adopted in 2016 lays down comprehensive guidelines for collecting, storing, and using personal data. While it is mainly aimed at protecting data, Article 22 speaks about "Automated individual decision making, including profiling", specifying that "*data subject 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*" (subject to exceptions for

contract enforcement, law and consent). It calls for consent, safeguarding the rights and freedoms, and further gives the subject the right to obtain human intervention, express their point of view and contest the decision (EU GDPR, 2016).

(Goodman & Flaxman, 2017) in their review of Article 22 of the GDPR reflect that this could necessitate a 'complete overhaul of widely used algorithmic techniques'. They look at this provision as a 'right to non-discrimination' and a 'right to explanation' when read with other articles in the GDPR. Contrary to this, (Wachter, Mittelstadt, & Floridi, 2016) argue that while the 'right to explanation' is viewed as an ideal mechanism to enhance the accountability and transparency of automated decision-making, there is doubt about the legal existence and feasibility of such a right in the GDPR, owing to the lack of explicit, well-defined rights and imprecise language. They contest that Articles 13-15 of the GDPR merely mandates that data subjects receive 'meaningfully, but properly limited information', what they call the 'right to be informed'. They raise the need for a meaningful right to explanation to be added to Article 22, where data controllers need to give the rationale for decisions, evidence for the weighing of features and logic of decision making.

In the Indian context, (Kapur & Khosla, 2019) observe that dealing with new technologies is one of the most demanding challenges facing regulatory design. (Padmanabhan & Rastogi, 2019) identify that the point of threat to individual and group rights has shifted from data gathering to data processing, and that the regulation of algorithms is unaddressed. Further, they note that there are no clear substantive safeguards against potential harm to social and individual rights, or regulatory mechanisms to mitigate against them in India.

The regulations or governance of algorithms could be cross-sectoral or/and sector-specific. A cross sectoral algorithmic governance could imply having a special regulatory or supervisory agency to audit algorithms and oversee its functioning. Calls have been made to establish for a FDA for Algorithms (Tutt, 2016), Machine Intelligence committee (Mulgan, 2016), an AI Watchdog (Sample, 2017), or a Algorithmic Safety Board akin to the US National Transportation

Safety Board (Shneiderman, 2017). Such bodies operating at a jurisdictional level would have the power to license algorithms, monitor their use, and investigate them. In addition to cross-sectoral regulations (or in the absence of it), sector-specific algorithmic regulations could operate. Regulators in different sectors like healthcare, finance or education can design rules and oversee the working of algorithms that operate in their sector. Given sector-specific challenges and algorithmic use cases, an overarching regulator might not have the capacity, time or domain-knowledge to address the issues (Andrews, 2017), and it may be inappropriate to solely apply solutions across sectors (New & Castro, 2018).

A crucial debate on the regulations is about the capacity of the regulators to deal with the ever-evolving nature and growing ubiquity of technology. The use of technology and algorithms are cutting across sectors and are increasingly used in finance, health, education, mobility, and more. To regulate rapidly transforming sectors, there has been a growing call for the use of RegTech. RegTech (or regulatory technology) are 'technological solutions to regulatory problems' (Chazot, 2015) that use technology for regulatory monitoring, reporting and compliance (Arner, Barberis, & Buckley, 2016). RegTech can use various technical, mathematical and statistical functions to detect financial fraud, biased practices, anti-trust activity, etc. It can also be implemented as tools using which regulators get automated compliance reports, allowing them to monitor tech without understanding its full working, enable cost savings and gain superior monitoring ability. (Arner, Barberis, & Buckley, 2016) refer to this as "the early signs of real-time and proportionate regulatory regimes."

SEBI guidelines for robo advisory tools

While there are no overarching regulations on algorithms in India; some sectoral regulators have delineated guidelines and regulations on use of algorithms in their sectors. Automated tools used in fintech are subject to regulations by the Securities Exchange Board of India (SEBI), a statutory body that regulates the securities market in India. In 2016, they released a consultation paper in which they lay down rules for 'Online Investment Advisory and automated tools' (SEBI, Consultation Paper on Amendments/Clarifications to the SEBI (Investment

Advisers) Regulations, 2013, 2016). In this section, they clearly state that automated tools need to comply with all rules under the SEBI (Investment Advisers) Regulations, 2013, over and above which they are subject to additional compliances.

One primary function of an investment advisor under the Investment Advisers Regulations is to profile the risk class of the user. The Investment Advisers regulations states that, "*Risk profiling of investor is mandatory, and all investments on which investment advice is provided shall be appropriate to the risk profile of the client*" (SEBI, SEBI (Investment Advisers) Regulations 2013 [Last amended on December 08, 2016], 2016). Further, it also says that the tools need to be fit for risk profiling and the limitations should be identified and mitigated. There are further rules that require them to act in the best interests of the client (i.e., the user of the tool), disclose conflicts of interest, and store data on the investment advice given.

Under the specific rules for automated investment advisory tools, firms are required to have robust systems and controls to ensure that any advice made using the tool is suitable and in the best interest of the user. They need to disclose to the user how the tool works and the limitations of the outputs it generates. The tools must undergo a comprehensive system audit and be subject to audit and inspection. Finally, regulations also mandate that robo advisory firms need to submit a detailed record of their process to SEBI. This includes the firm's process of risk profiling of the user and their assessment of the suitability of advice given, which is to be maintained by the investment adviser for a period of five years.

Explainable Algorithmic Decision Systems (ADS)

Algorithms are 'black boxes' and users affected by it know little to nothing about how decisions are made. Being transparent and explaining the process helps build trust in the system and allows regulators and users to hold it accountable. With their growing ubiquity and potential impact on businesses, 'explainable AI' (xAI) or more generally, 'explainable algorithmic decision systems' is more necessary than ever.

Explainability has been defined in various ways in research. The most prominent one, given by FAT-ML considers an algorithm explainable when it can "*Ensure that algorithmic decisions as well as any data driving those decisions can be explained to end-users and other stakeholders in non-technical terms*" (Diakopoulos, et al.). They identify 'explainability' as one of the five principles for accountable algorithms. The other four are responsibility, accuracy, auditability, and fairness.

The literature on explainable ADS is vast and is constantly growing. This section covers literature on the ways in which the models can be explained, the types of models that can be explained, and the trade-offs to explanations.

(Castelluccia & Le Métayer, March 2019) in their report identify three approaches to explainability. A 'black box approach', 'white box approach' and a 'constructive approach'. The black box approach attempts to explain the algorithm without access to its code. In this approach, explanations are found by observing the relationship between the inputs and outputs. In the white box approach, the code is available and can be studied to explain the decision-making process. The constructive approach is a bottom-up approach that keeps explainability in mind before and while coding the ADS, thus building in 'explainability by design'.

Explainability is affected by the type of algorithm as well. While some models are easy to explain with or without access to the code, complex ML and neural network models are very difficult to explain to humans. Explainability is easier in parametric methods like linear models where feature contributions, effects, and relationships can be easily visualised and the contribution to a model's overall fit can be evaluated with variance decomposition techniques (Ciocca & Biancotti, 2018). However, that task becomes tougher with non-parametric methods like support vector machines and Gaussian processes and especially challenging in ensemble methods like random forest models. For example, in fintech, an ML model used to predict loan defaults may consist of hundreds of large decision trees deployed in parallel, making it difficult to summarize how the model works intuitively (Bracke, Datta, Jung, & Sen, 2019). The newer methods of deep learning or neural networks pose the biggest challenge, they are able to model

complex interactions but are almost entirely uninterpretable as it involves a complex architecture with multiple layers (Thelisson, Padh, & Celis, 2017) (Goodman & Flaxman, 2017). Currently, there is a significant academic effort in trying to demystify these models. As it gets increasingly complex, there is also a call to avoid altogether using uninterpretable models because of their potential adverse effects for high stakes decisions, and preferably use interpretable models instead (Rudin, 2019). Several explainability methods for parametric and non-parametric models have been researched for this paper, and have been briefly covered in the methodology section.

The quality of explanations is evaluated by several indicators such as their intelligibility, accuracy, precision, completeness and consistency. There can often be a trade-off between them. By focussing on completeness, the intelligibility of the explanation can be compromised.

Research statement

Our research helps explain how an algorithm-based decision-making “black box” works, specifically in determining the risk profile of users in robo financial advisory apps. For this, we propose a RegTech tool to explain the algorithms decision making to regulators.

Research objective

Building user trust, especially in matters of personal wealth investment, would increase engagement with robo investment advisory services and allow users to reap the benefits they offer. Giving ‘explanations’ that describe the decision-making process and the parameters used for it is one way through which trust can be built. Additionally, explanations promote transparency and open it up to regulatory oversight. A regulator auditing the algorithm-based on a set of pre-defined regulations or guidelines would increase user trust and ensures that automated investment advisors are unbiased, acting in the best interests of the user, and do not face a conflict of interest. With comprehensive and meaningful explanations, regulators could audit the algorithms and check if they comply with the regulations they are subject to.

Algorithms used in automated wealth or investment advisory tools are subjected to SEBI regulations in India. However, regulators without the technical knowledge possess no means to understand the algorithms and test it themselves. The objective of our research is to develop a **RegTech** tool with customised explanations that can be used by regulators to understand and evaluate the decision making of any robo-advisory application ADS.

Research Questions

1. What methods from xAI can we use to operationalise explainability in the risk-category profiling done by a robo financial advisor algorithm?
2. Can the process of algorithmic explainability be standardised for regulators and for different data types and algorithms?
3. To what extent can these methods be used to satisfy the regulatory requirements that robo financial advisors need to comply with?

To design a study that can explain the questionnaire-based risk profiling done by robo advisors, the boundaries of the study have to be defined. Four methodological considerations have been discussed in this section.

Defining the boundaries of the study

The first consideration for operationalising is deciding **the depth of review/assessment** by looking at the decision-making process; this depends on the availability of required inputs for assessment. As mentioned, there is a white box and a black box approach. For the white box approach, it is essential to know how the computer makes decisions. This necessitates the third party assessing the algorithm to be given access to the algorithm. While this would greatly aid transparency, they are the intellectual property and trade secrets of the robo advisory firms. This is also the case for robo financial advisory apps. Thus, in the absence of the code, the second "black box" approach is used. Given the "black box" nature of algorithms, alternate methods are used to check if inputs give the intended outputs, to check the representation of training data, identify the importance of features, and find the relationship between them. Robo financial advisors would not disclose their code or algorithm used for decision making, and hence, we will use black box explainable methods. The firm would have to provide

a dataset with a sample of its input criteria and corresponding predicted output to the regulator (i.e., the input-output data).

Second, there is a limitation to the **level of simplification** of a black box algorithm. As mentioned, there is a trade-off between complexity, completeness, and accuracy of the system and its explainability. The RegTech tool does not have access and thus does not know the algorithm used by the robo-advisor — it could be simple parametric models, the more complex non parametric models or neural networks. Our study is limited to developing a tool that can explain parametric and non-parametric models. To do this, we will employ methods from Machine Learning. Neural networks have not been tested and is not in the scope of this study.

Third, we have **global and local explanations**. Global methods aim to understand the inputs and their entire modelled relationship with the prediction target or the output. It considers concepts such as feature importance, or a more complex result, such as the pairwise feature interaction strengths (Hall & Gill, 2018). Most feature summary statistics can also be visualised by using partial dependence plots or individual conditional plots. Local explanations in the context of model interpretability try to answer questions regarding specific predictions; why was that particular prediction made? What were the influences of different features while making that specific prediction? The use of local model interpretation has gained increasing importance in domains that require a lot of trust like medicine or finances. Given the independent and complimentary value added by both methods, we will include both global and locally interpretable explanations in our study.

Finally, there is a challenge in **communicating** the results. This depends mainly on the end-user— the person who will view the explanation report. The report would have to be designed based on why they want to see the findings, and what their technical capability, statistical, and domain knowledge is. If the end-user is a layperson wanting to understand how the algorithm makes decisions at a broad level, the tool would need to be explained in a very simplified and concise manner. In contrast, if the end-user is a domain expert or a regulator who is interested in

understanding the details and verifying it, the findings reported would have to reflect that detail. As mentioned in the objectives, the end user for our explanation report is a regulator. In addition to this, a branch of study called Human-Computer Interface (HCI) focuses specifically on choosing the best communication and visualization tools. Our study does not focus on this aspect, but rather confines itself to employing appropriate explainable methods for a regulator.

Hence, our tool narrows the scope of the study to the following — explaining the robo advisors black box algorithm by approximating a best fit model to a given data set. Followed by explaining the trends and decisions observed in the dataset using global and local explanation methods. These explanations will be aimed at the regulator.

Methodology

The research aims to explain the questionnaire-based risk profiling done by any robo advisor using algorithms to reverse engineer key aspects of the decision making. **To study this in the absence of the algorithm, firms will have to provide regulators with the questionnaire, a sample of the user responses (input criteria) and the corresponding risk category predicted by the algorithm (i.e., the input-output data).** (Part 1 of the findings quantifies the sample size that needs to be provided). Using this RegTech tool, regulators will be able to audit the algorithm to check if it complies with the regulations.

The methodology details how the tool reverse engineers the input-output data in order to understand how the algorithm takes a decision and is divided into three parts.

The first part talks about how the sample dataset required for the study was generated. In the absence of real-world data, a sample representative dataset had to be generated on which the explanation methods could be tested. To ensure that the results of the study are replicable for any type of equation used by the algorithm, we used several different methods to generate this sample dataset.

The second part looks at the information that the tool RegTech tool needs to provide to explain the robo advisors ADS to the regulator. Information about how much each response contributes to the decision and how they relate with each other have been addressed in this section. Three explanation methods have been identified (two global and one local explanation).

In the third and final section, the technical aspects of three explanation methods for the robo advisory ADS has been detailed.

Before proceeding, we clarify the meaning of three terms that are commonly used in machine learning and data analysis, and explain what they mean in the context of our study (see diagram in [Appendix 1](#))

- Each question in the questionnaire is a 'feature' in the dataset. The 'weights' associated with each feature contributes to the decision made by the ADS.
- 'Categories' refers to the options for a question (or equivalently, the response given to a question)
- The risk classes ('no risk' to 'high risk') that robo advisors assign to a user are 'classes'. There are 5 classes in this study.

That is, each question (feature) in the questionnaire has options (categories). Based on responses users can give, they are assigned a score. The output generated after answering all the questions in one out of five risk class, ranging from 'no risk' to 'high risk'.

Other definitions and terms from machine learning and statistics that have been used in the methodology and findings are explained in [Appendix 1](#).

1. Generating the dataset for the study

To conduct this study, we needed to generate a sample data set that can adequately represent the real world. The reliability would have to be such that it can work for input-output data from any robo advisory app. In other words, the analysis should be able to handle any number of questions, any type of question (i.e., questions with categorical or continuous variables as its options), and any number of options. Additionally, a controlled generation of dataset allows us to build in some trends in the data. If the explanations can accurately capture these trends without access to the equations used to generate it, then we can conclude that the explanations are successful in accurately reverse engineering the decision making of the algorithm.

For our study, we surveyed several robo advisors and used the questionnaire from Paytm Money to create a data set with all possible user profiles. All major robo financial advisors were surveyed at the time of writing this paper. The reason Paytm Money was used to create the said data set is because it is one of the leaders in terms of market penetration. Also, other robo advisory applications use

similar questions, and therefore, the choice of questionnaire is not of great significance.

Step 1- The robo advisory questionnaire is used to model an equation by giving weights to each question (i.e., feature). It is converted to a format such that output is a function of the features and weights. The equation can be represented as follows-

$$\text{output} = f(w_1x_1, w_2x_2, w_3x_3 \dots w_nx_n)$$

where x_i represents the response to question 1 and w_i is any real number that represents the weight given to question 1. 'f' is the function that models their relationship. For example, if the questionnaire has two questions and question 1 is about the age of the respondent and question 2 about the salary of the respondent, the output risk category could be modelled by an equation like: *risk category* = $w_1(\text{age}) + w_2(\text{salary})$.

Step 2- A score is assigned to each option ('category') in each question. For example, within the question about age, the option with age group 18-35 could have a category score of 1, age-group 36-55 a score of 0.5 and so on. For our study, the scores assigned to each category is given in [Appendix 2](#). The scores we have used are only indicative and have no significance. Appendix 2 explains how the features are ranked. It is important to note that the tool is valid for any input equation with any score.

Step 3- Using the questions (i.e., features) and options (i.e., categories), all possible combinations of user profiles are generated. A stratified sample of the dataset is taken for further analysis. This is equivalent to the 'input' part of the input-output dataset that the firm need to provide to the regulator.

Step 4- Using the values from step 1 and step 2, the output score is calculated for every user profile. The entire range of scores is divided into five risk classes in order to put each user in one of five output classes — no risk, low risk, medium risk, likes risk, and high risk. These classes are the 'output' part of the input-output dataset that the firm need to provide to the regulator.

The firm needs to provide a sample of the inputs and corresponding outputs to the regulator. The detailed process, equations used for this study and profile of the selected dataset can be found in [Appendix 2](#).

Validity and reliability checks-

- In order to ensure that the dataset is an accurate representation of reality, data from PayTM was used. Because the process we use is independent of the number or type of features and categories, it can be replicated for any robo-advisory application.
- In order to ensure replicability, robustness and reliability of results, in step1, several types of input models were used. For our study, we tested four types of possible algorithmic equation types that could be used to generate datasets- a linear equation under independent variable assumption, an equation with interaction effects, quadratic and logarithmic generation. The details of the equations and sample is given in Appendix 2. The results for all types of equations have been reported in the findings.
- The process we use is also independent of the score associated with options (step 2). Hence, the study is valid for all values.

2. Information that needs to be explained by the robo-advisory

To explain the internal mechanics and technical aspects of an ADS in non-technical terms, we need to first identify the instances of decision-making which are opaque in order to make them transparent and explain them.

Robo advisors conduct a complex assessment of the users age, financial condition, risk tolerance, capacity, and more, to classify the user in to a risk class, and use it to offer automated advice suited to their investment goals. There is no way to ascertain that the advice given is not unwittingly biased, has unintended correlations or is giving undue importance to certain undesirable features (for example, the Apple credit card was accused of reproducing a gender bias because the algorithm gave a 20 times higher credit limit to a man as compared to his wife;

both with the same financial background (Wired.com, 2019)). Thus, there is a need to explain the rationale for the risk classification and show that there is no undesirable importance given to certain features. In practice, if the robo advisor asks questions on age and salary, the explanation would need to tell which of the two features is more important and by how much. If gender is one of the input parameters, the explanations would be able to tell if that particular question has an undue influence on the output. Apart from this, we also need to give the regulators the ability to spot check the output. For any randomly selected user profile, a "local" explanation will allow the regulators to understand how the algorithm processes one data point and if the generated output aligns with the expected output.

In our study, we generate three explanations (two global and one local) that the regulator can use to understand how the robo advisor takes a decision.

- **Feature importance scores:** this provides a score that indicates how useful or valuable each feature (i.e., question) is in the construction of the model. If the weightage given to a feature is large or if the feature is higher up in a decision tree algorithm, it has a higher relative importance. In our case, feature importance scores will tell us the relative importance of the features and their contribution to the risk classification.
- **Feature relations:** this tells us how features relate to each other and with the output. insights can be gained by examining the behaviour of different categories (options) within each feature (question) and how they vary with each other and affect the output. In our case, we can use these methods to find the relationships between the features, its categories and the output risk classes that are built in the black box algorithm.
- **Local explanations:** Local explanations in the context of model interpretability try to answer questions regarding specific predictions; why was a particular prediction made? What were the influences of different features while making that particular prediction? As mentioned above, in our case, local explanations will help explain why a particular

user was assigned a particular risk class. It can also be useful to understand boundary points and outliers.

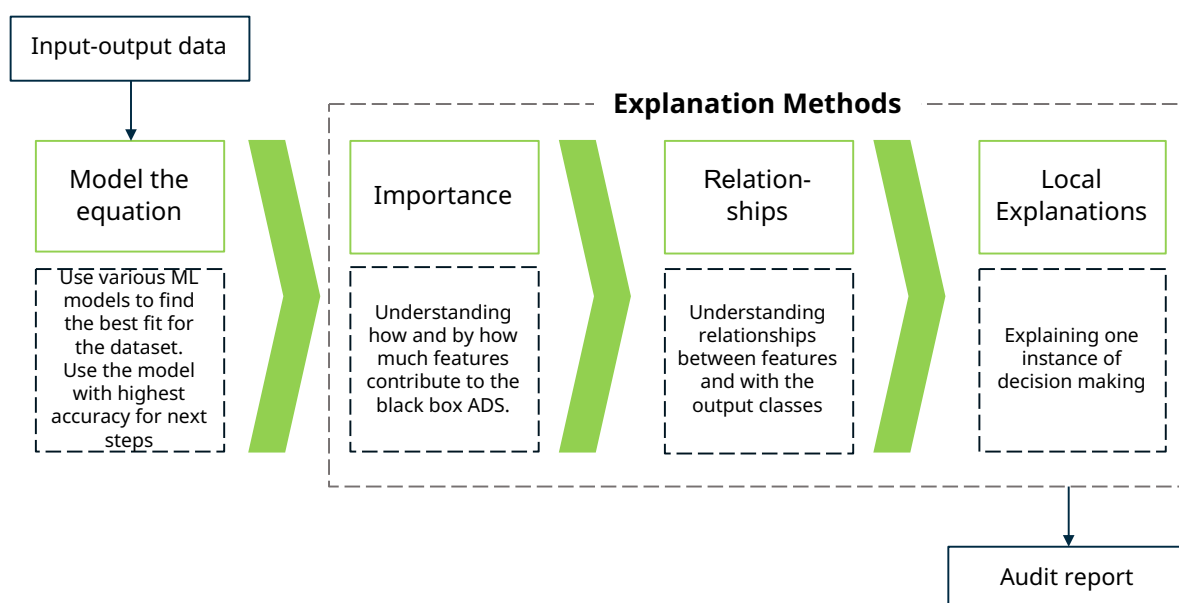


Figure 2. 1 Framework for explanation

Fig 2.1: Firms would have to provide a stratified balanced sample of the input-output data to the regulator. The RegTech tool will model the data to an equation and generate an audit report with three explanations.

3. Operationalising the explanations

In order to find the feature importance scores, feature relations and local explanations, we reviewed and tested the several methods that researchers have developed. Various toolkits have been developed to operationalise concept of fairness, accountability, transparency and explainability in algorithms. In our review of tools, we found FairML, LIME, Aequitas, DeepLift, SHAP and VINE to be particularly relevant. Most of the toolkits focused on explainability, while only a handful try to operationalise fairness. While toolkits like FairML and LIME aim to be a generalized method or tool that is sector agnostic, others have developed techniques to address the domain-specific issues (for eg. DeepLift is used for genome sequencing). Consequently, the end product of the two approaches varies between easily understandable by all to interpretable only by domain experts. We also explored the viability of statistical methods like LASSO (least absolute shrinkage and selection operator), mRMR (minimum Redundancy Maximum Relevance) feature selection and random forest algorithms.

Firms would have to provide a stratified balanced sample of the input-output data to the regulator. The RegTech tool will model the data to an equation and generate an audit report with three explanations described above.

The first step is that of modelling the input-output data to an equation with no information about the method or logic used by the firm to arrive at the decision. We do this using machine learning models. Following this, the three explanations (feature importance scores, feature relations, and local explanations) are generated.

3.1. Modelling the dataset accurately

To model the input-output data, five **supervised ML models** are used. Firms provide the regulator a stratified sample of the input-output data. The dataset classifies inputs in five output classes (high risk to no risk) making this a multiclass type of classification. This sample is divided into two parts, the 'training data' and the 'test data'. The training data is used to train the ML models. The models then try to predict the outputs from the inputs in the test data, checks if the predicted output and the actual output match and determines the accuracy of the fit. This is repeated for multiple types of input equations to check for the reliability of the models. Overfitting is not a worry here as we are not using the model to predict new data, rather the aim of fitting a model here is to give us a better representation of the data set, and a higher accuracy indicates that the ML model is able to better reflect reality.

A variety of classifiers are available to model these mapping functions. Each ML classifier adopts a hypothesis to identify the final function that best fits the relationship between the features and output class. The input-output dataset was modelled using five machine learning algorithms frequently used for predictions; logistic regression (LR), support vector machines (SVM), decision trees (DT), naive bayes (NB) and k-nearest neighbours (KNN). These algorithms were chosen based on difference in 'hypothesis functions' and each model is good at recognising different feature relationships and interactions. The explanation of these models and how they work can be found in [Appendix 3](#).

The ability of the model to accurately describe the dataset is given by commonly used performance measures such as accuracy, precision, recall, and the f1 score. The definitions are given in [Appendix 1](#). The five models are run and the model that performs best based on these metrics are selected for further explanations.

3.2. Finding Feature importance scores using shapley values

As mentioned, feature importance scores give the relative contributions made by each feature (question) in the risk classification decisions made by the ADS. To find these contributions we use the concept of shapley values, commonly used to decide relative contributions made by each feature in game theory. This is generated from the SHAP library, a unified framework built on top of several model interpretability algorithms such as LIME and DeepLIFT. The SHAP package can be used for multiple kinds of models like trees or deep neural networks as well as different kinds of data including tabular data and image data.

If we have 3 features (A,B,C) contributing to the output of the model then these features are permuted (B,C,A or C,A,B, etc..) to give new target values that are compared to the originals to find an error. Thus shapley values of a feature are its average marginal contributions across permutations. Shapely values are relative, thus the impacts made by each feature makes sense only in the context of other features. This means as the features/ questions change we will see different patterns emerging.

3.3. Determining feature relations using partial dependence plots

Once the important features are identified, we need to assess the interactions and relationship between them (or a subset) and the response. This can be done in many ways, but in machine learning it is often accomplished by constructing partial dependence plots (PDPs), and we use this method in our study. These plots portray the marginal effect one or two features have on the output risk classes and visualises the relationship.

PDP can be used as a **model agnostic global level understanding** method to gather insights into black box models. Model agnostic means that PDP's make no

assumptions regarding the underlying model. The partial dependence function for regression is defined as-

$$f_{x_s}(x_s) = E_{x_c}[f(x_s, x_c)] = \int f(x_s, x_c)dP(x_c)$$

x_s is the set of features we find interesting, x_c is the complement of that set (set of all features we don't find interesting but are present in the dataset), $f(x_s)$ gives the partial dependence and $P(x_c)$ is the marginal probability density of x_c . f is the prediction function.

The whole function $f(x_s)$ is estimated as we don't know the f (it's model agnostic) nor do we know the marginal probability distribution.

$$f_s = \frac{1}{N} \sum_{i=1}^N f(x_s, x_{c_i})$$

The approximation here is twofold: we estimate the true model with f , the output of a statistical learning algorithm, and we estimate the integral over x_c by averaging over the N x_c values observed in the training set.

3.4. Local explanations

Local explanations mean explaining a single instance of decision made by an ADS system. To find the logic behind these decisions we used LIME, or Locally Interpretable Model-agnostic Explanations. This method, developed by a group of researchers, uses local surrogate models to approximate the predictions of the underlying black box model. Local surrogate models are interpretable models like Linear Regression or a Decision Trees that are used to explain individual predictions of a black box model (Ribeiro, Singh, & Guestrin, 2016). LIME trains a surrogate model by generating a new data-set out of the datapoint of interest. The way it generates the data-set varies dependent on the type of data. For text and image data LIME generates the data-set by randomly turning single words or pixels on or off. In the case of tabular data, LIME creates new samples by permuting each feature individually. The model learned by LIME generally is a good local approximation of the black box model and gave satisfactory results for our study.

Findings

The findings are divided in four parts. The first part gives the results of the Machine Learning models that are used to fit the input-output data and reverse engineer the importance of features in the robo advisors risk profiling. The second and third parts explain the risk profiling using global explanation methods. The second part reports the feature importance scores and the third part reports the feature relations. The fourth and final part of the findings provides the local explanations to spot-check the algorithm or explain one specific decision made by it.

The findings have been reported for divergent cases, which are representative of the overall findings. All test cases, generation of sample data, and findings can be accessed through [this GitHub link](#).

Part 1- modelling the risk profiling decision

The aim of the first part is to fit a model to the input-output data that can predict the outputs as accurately as possible. As mentioned in the methodology, in this step, various ML models are used and the most accurate model is identified. This first step is crucial because the best-fit model is required to implement the three explanation methods.

The accuracy of the prediction and the f_1 -scores of the classes need to be considered together to select the best model for the dataset. The results (for five ML models and four input equations) have been summarised in the table below (Table 2.1).

Table 2. 1 Summary of results for the ML models and input equations

	Linear Equation under independent variable assumption		Quadratic Equation		Equations with interaction effects		Logarithmic Equation	
Performance Metrics	Accuracy (%)	F1 - Score	Accuracy (%)	F1 - Score	Accuracy (%)	F1 - Score	Accuracy (%)	F1 - Score
Logistic Regression (LR)	90	- no risk : 0.49 - low risk : 0.91 - moderate : 0.93 - likes risk : 0.86 - high risk : 0.52	78	- no risk : 0.88 - low risk : 0.77 - moderate : 0.73 - likes risk : 0.71 - high risk : 0.91	78	- no risk : 0.96 - low risk : 0.79 - moderate : 0.23 - likes risk : 0.80 - high risk : 0.94	76	- no risk : 0.91 - low risk : 0.77 - moderate : 0.00 - likes risk : 0.78 - high risk : 0.94
Gaussian Naive Bayes (GNB)	75	- no risk : 0.56 - low risk : 0.71 - moderate : 0.81 - likes risk : 0.65 - high risk : 0.26	70	- no risk : 0.79 - low risk : 0.76 - moderate : 0.68 - likes risk : 0.51 - high risk : 0.72	67	- no risk : 0.95 - low risk : 0.82 - moderate : 0.43 - likes risk : 0.00 - high risk : 0.40	68	- no risk : 0.77 - low risk : 0.61 - moderate : 0.58 - likes risk : 0.64 - high risk : 0.80
K- Nearest Neighbours (KNN)	93	- no risk : 0.90 - low risk : 0.94 - moderate : 0.94 - likes risk : 0.92 - high risk : 0.86	96	- no risk : 0.97 - low risk : 0.96 - moderate : 0.96 - likes risk : 0.95 - high risk : 0.96	97	- no risk : 0.99 - low risk : 0.98 - moderate : 0.96 - likes risk : 0.95 - high risk : 0.94	98	- no risk : 0.98 - low risk : 0.97 - moderate : 0.96 - likes risk : 0.98 - high risk : 0.99

Support Vector Machines (SVM)	98	- no risk : 0.63 - low risk : 0.97 - moderate : 0.99 - likes risk : 0.99 - high risk : 0.94	93	- no risk : 0.94 - low risk : 0.92 - moderate : 0.93 - likes risk : 0.93 - high risk : 0.94	96	- no risk : 0.96 - low risk : 0.95 - moderate : 0.95 - likes risk : 0.95 - high risk : 0.95	92	- no risk : 0.92 - low risk : 0.89 - moderate : 0.87 - likes risk : 0.94 - high risk : 0.96
Decision Trees (DT)	89	- no risk : 0.81 - low risk : 0.89 - moderate : 0.90 - likes risk : 0.88 - high risk : 0.81	96	- no risk : 0.97 - low risk : 0.95 - moderate : 0.95 - likes risk : 0.95 - high risk : 0.95	98	- no risk : 0.99 - low risk : 0.98 - moderate : 0.97 - likes risk : 0.95 - high risk : 0.94	99	- no risk : 0.99 - low risk : 0.99 - moderate : 0.99 - likes risk : 0.99 - high risk : 0.99
Best Model	K - Nearest Neighbours		K - Nearest Neighbours		Decision Tree		Decision Tree	
Explanation	K - Nearest Neighbours can generate a highly convoluted decision boundary, hence points that are very close to each other can be modelled very well using this method				DTs perform very well for all input equations except the linear model. It gives very accurate results because the options are categorical, which DT can identify much better			

Table 2.1- accuracy of the prediction and the f1-scores of the classes for five models (LR, GNB, KNN, SVM, DT), and four input equations (linear equations under independent variable assumption, quadratic, equations with interaction effects and logarithmic)

As our findings show, KNN fits linear (under independent variable assumption) and quadratic equations most accurately and the Decision Tree model fits equations with interaction effect and logarithmic equations most accurately. The findings also highlight why it is not sufficient to consider only the accuracy. Take the example of SVM on a linear equation. It gives a high accuracy of 98%, higher than the KNN model. However, the f1 score of the no-risk class is only 0.63. This indicates that the SVM model can make very good predictions for other classes, but fails to do it in the no-risk class.

The RegTech tool will run the sample input-output data provided by the firm. The four ML models will model it. The model that maximizes accuracy and f1-score will be selected and used as a basis for generating the explanations.

Optimal size of input-output sample data that the RegTech tool requires

What is the minimum size of training data that firms should share with the regulator without compromising the accuracy of the modelling? While there are thumb rules and more is considered better, we report the minimum sample required. To find this, we ran the models with different sample sizes in order to provide a ball-park figure or the number of data points that need to be provided by the robo advisory firm to the regulator.

Stratified samples of the input-output data of different sizes were selected as the training data, the ML models were run on them and the accuracy and f1 scores were found. The sample sizes included values between 1.5% of the training data to 12% of the training data.

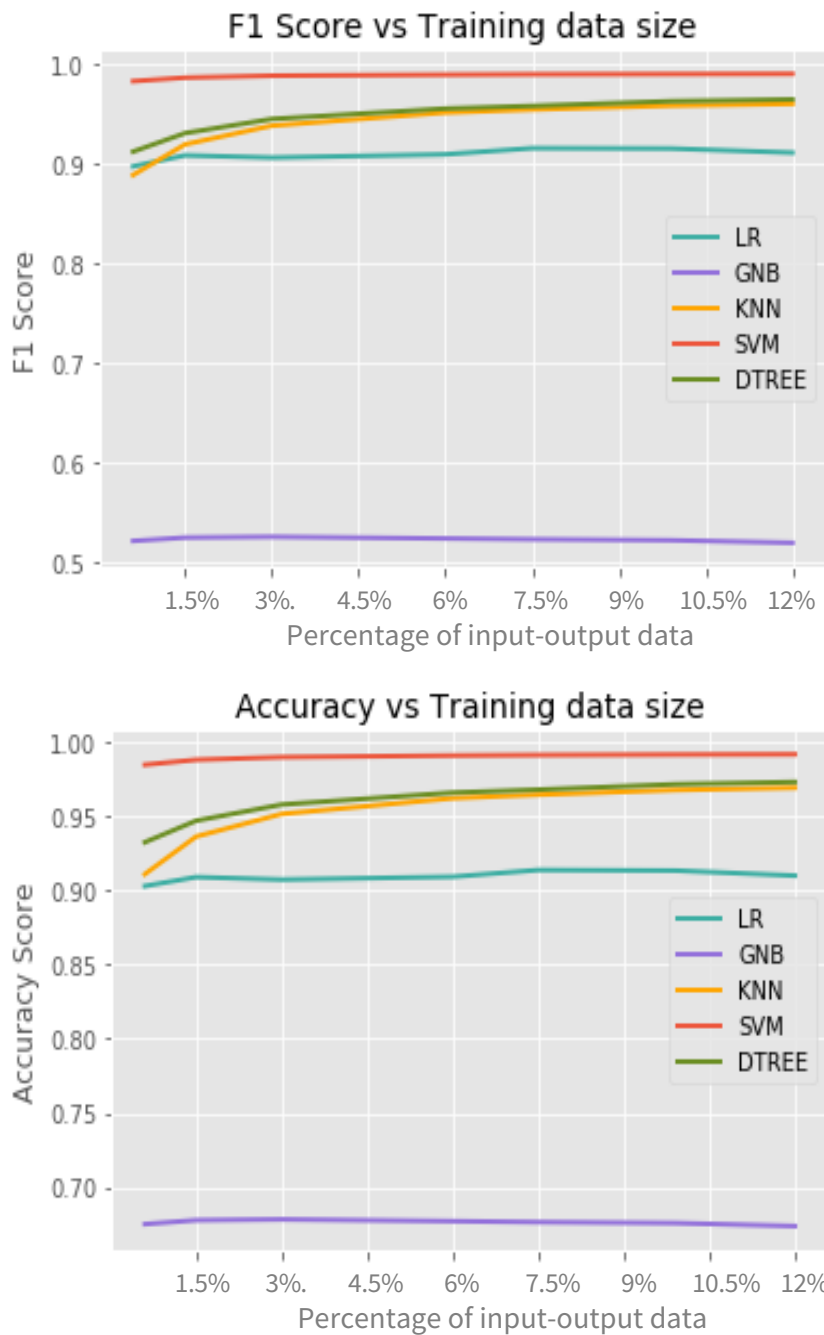


Figure 2. 2 F1 scores and accuracy versus size of training data

Figure 2.2: Graph on top - F1 scores (y-axis) versus percentage of input-output data used as training data (y-axis). Graph at bottom - Accuracy (y-axis) versus percentage of input-output data used as training data (y-axis).

The primary goal of these graphs is to determine whether a change in sample size affects the F1 Score or accuracy level, in order to ascertain the optimal sample size for analysis. Lines of different colours represent the results for different ML models.

LR- linear regression; GNB- Gaussian Naïve Bayes; KNN- K-Nearest Neighbours; SVM-Support Vector Machines; DTREE- Decision Tree

As expected, the accuracy is directly proportional to the sample size: considering a larger sample gives greater accuracy. However, findings show that the relationship is not linear. The accuracy of most models increases with the increase in sample size till about 6% of the data and then stabilises. Amongst all the models, SVM performs the best with all sizes of data, followed by KNN and the DT model.

A 6% stratified sample of all input-output data, which translates to 67500 data points, would be sufficient in our case to run these models. Therefore, firms would need to give the regulators a minimum of 6% of the training data or ~67,500 data points, whichever is higher.

Part 2- Feature importance scores

Feature importance scores are part of the global explanations and have been found using the SHAP values. They have been represented using SHAP plots. They tell how and by how much each feature (question) contributes to the ADS risk classification process. We report two importance scores — the feature importance and the class-wise feature importance.

Importance scores for all equation types used to [generate the dataset for the study](#) were calculated. However, in the following sub-sections of the findings, the results showcase the scores obtained for equations that have interaction effects. As previously stated, we surveyed many robo advisers and utilised the Paytm Money questionnaire to generate a data set with all potential customer profiles. Furthermore, numerous interdependencies within variables have been investigated by examining interaction effects within variables using different quadratic and polynomial equations.

It is important to remember that these explanation methods are replicable for any set on input features, including demographic features (like gender, race), behaviour (such as purchase history or internet activity) or opinions (like political leaning).

2.1. Feature importance-

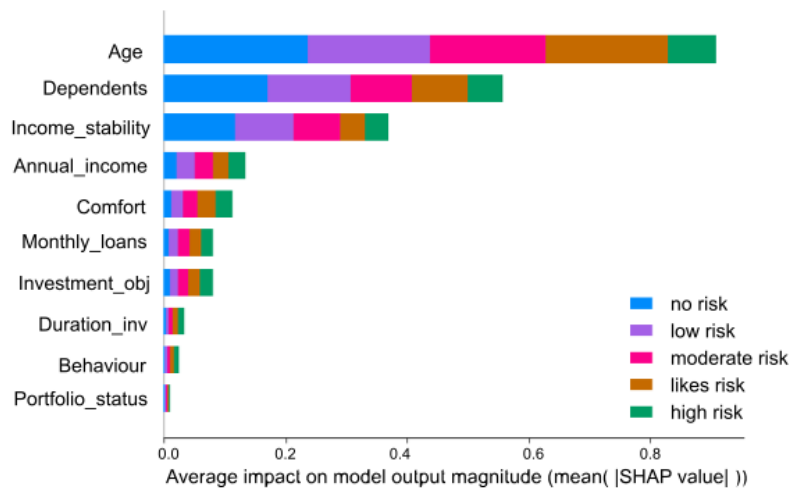


Figure 2.3 Overall feature impact on outcomes

Figure 2.3 shows the average impact of features on risk classes. The y-axis lists the features used to decide the risk class in descending order of importance. The x-axis shows the shapely value that quantifies the influence. The length of the bar indicates the total contribution of the feature to the output class. The colours indicate the average contribution of the feature different risk classes.

The SHAP plot shows that AGE is the most important feature while predicting the risk class for all output classes ('no risk' to 'high risk')

Figure 2.3 shows that 'Age' is the most important feature in the model, and has the greatest contribution to the risk categorization process. This accurately represents the weights that were given when the [dataset for the study was generated](#), indicating that the explanation method is successful in reverse engineering the input-output data without having access to the model. This will allow regulators to understand if any undesirable feature has a disproportionate importance score.

2.2. Class-wise feature importance

Feature importance scores help understand the importance of questions. Class-wise feature importance plots show how the categories in each feature behaves differently in different output class ('no risk' to 'high risk') and quantifies the effect. For instance, if a person has a large loan amount to repay every month,

their response should negatively contribute to the high risk class and positively contribute to the low risk class. Further, it shows the relative importance of features and the distribution of the stratified sample in the output class.

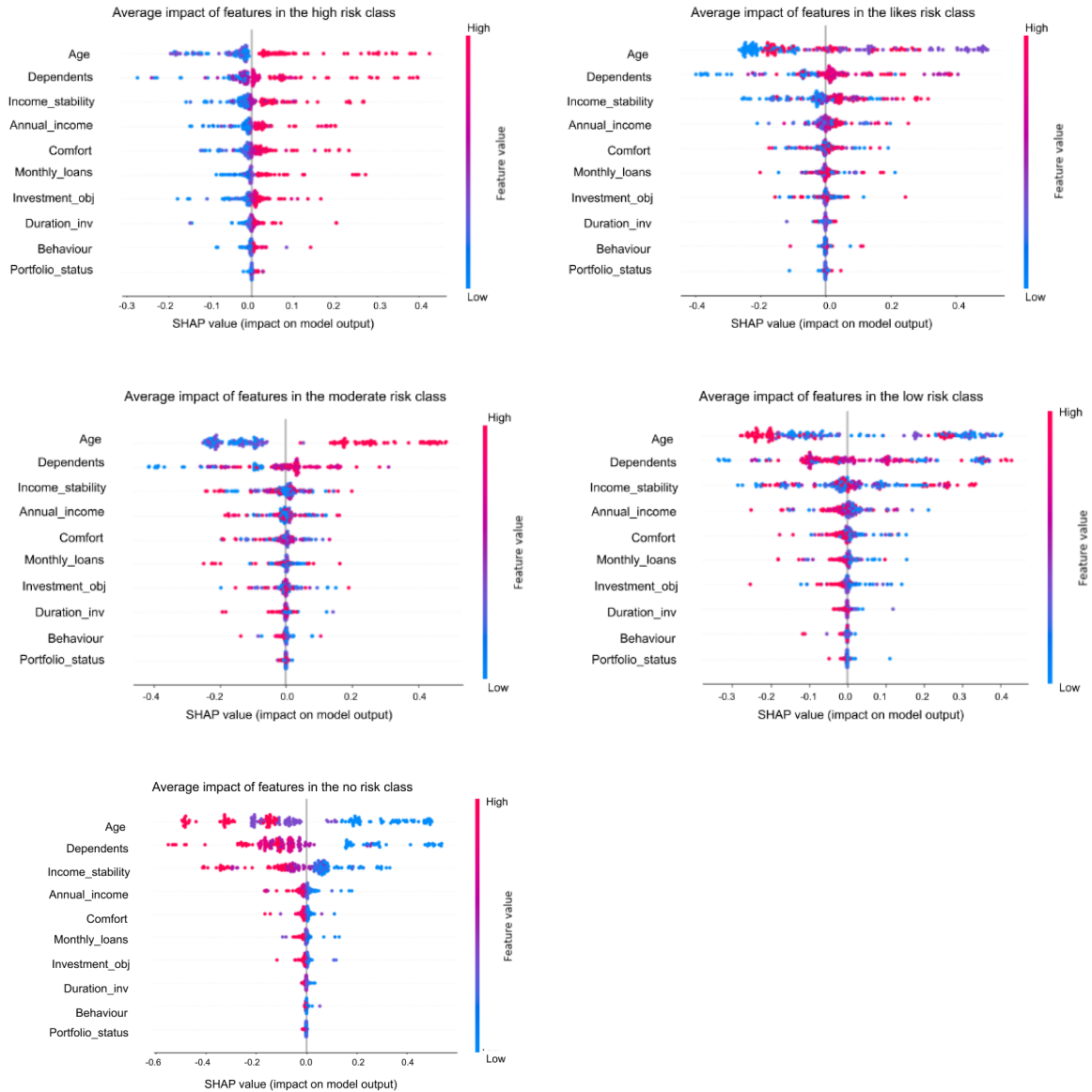


Figure 2.4 Class-wise feature importance plots

AGE SCORES:

18 - 35	36 - 55	55+
1	0.5	0.2

Figure 2.4 shows the class-wise feature importance plots for the five output risk classes ('no risk' to 'high risk'). The features are listed on the y-axis (in descending order of feature importance) and shapely values that quantify the effect are shown on the x axis. The colour represents the score of the

categories in a feature. The distribution of points in each feature represents the distribution of data points in the sample and also shows the extent of negative or positive influence.

The scores for the feature 'age of the user' that is used by the ADS is shown in the table. A person whose age falls in the 18-35 category is assigned a score of 1.

Using the 'age' feature as an example to interpret the graphs, it can be seen that a young person (in the age category of 18-35) has a high category score (of 1). Thus, according to the first graph in Figure 2.4, this demographic feature would result in a large positive contribution to the 'high risk' output (a positive shapely value of ~0.4). Similarly, an older person (age category of 55+ and a small category score of 0.2) will negatively contribute to the 'high risk' class. Additionally, features like the investment objective and monthly loans contribute very less to the extreme classes ('high risk' and 'no risk'), but influence the output significantly in the 'moderate risk' class. Once again, we observe that these plots can accurately represent the trends in the model without having access to it.

Hence, using this, regulators can understand the how various categories in a feature (for example gender being female) an affect an output, and by how much.

Part 3- Relationships

This section reports the relationship between features and the output class by showing how the changes in one or more feature changes the output.

One simple way of finding the relationships is to see the correlations between features and between features and the output, as shown in the correlation matrix in Fig 2.5.

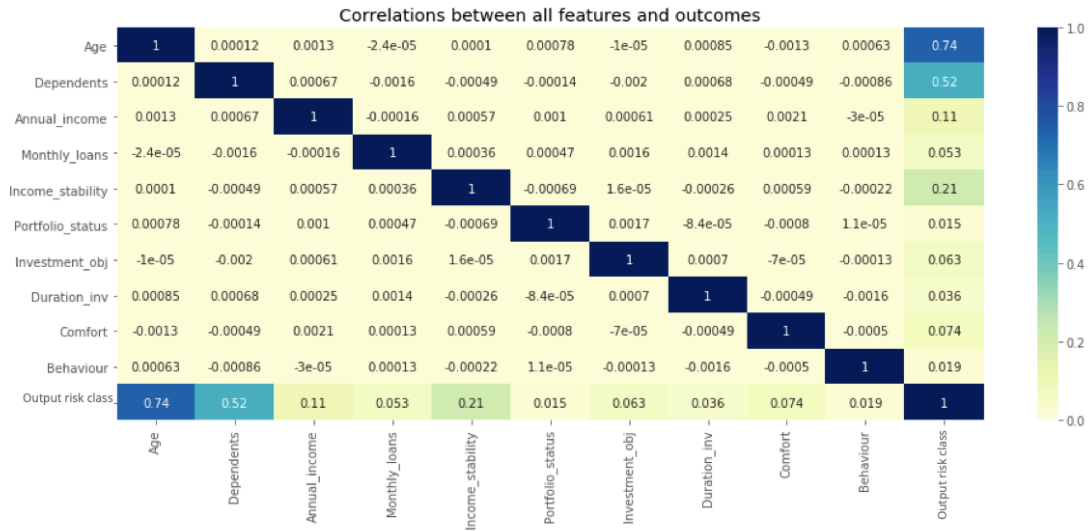


Figure 2.5 Correlation between features and outcomes

Fig 2.5: The correlation matrix shows the correlation coefficient between features and with the output class. A dark colour indicates a higher correlation.

Relationships can also be visualised using partial dependence plots between one feature and the output or two features and the output. Fig 2.6 shows the partial dependence relationships between one feature (age) and the output risk class decision.

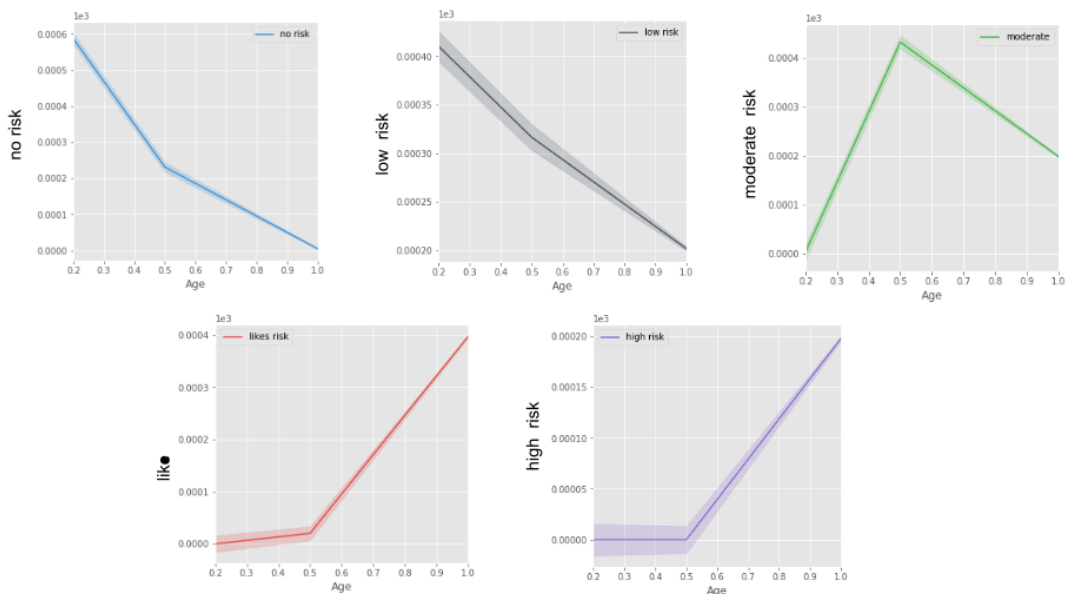


Figure 2.6 Relationship of age with output risk classes

Fig 2.6 shows the relationship of AGE with output risk classes ('high risk' to 'no risk'). As age increases (i.e., the age score decreases), the contribution to 'high

risk' class decreases. For the moderate risk class, there is an inflection point indicating that that very high or very low age scores would negatively contribute to the 'moderate risk' class.

Generation, analysis and sampling of data are done using logarithmic relationships in the polynomial equation. The dataset includes 37,50,000 entries for each of the 11 variables.

This shows how the changing scores of categories in a feature relate with the output class. It helps visualise the shape and the inflection point of the relationship, allowing the regulator to identify breaks where the effect of a feature could change drastically.

Similar partial dependence plots can be drawn to identify the relationship between two features and the output.

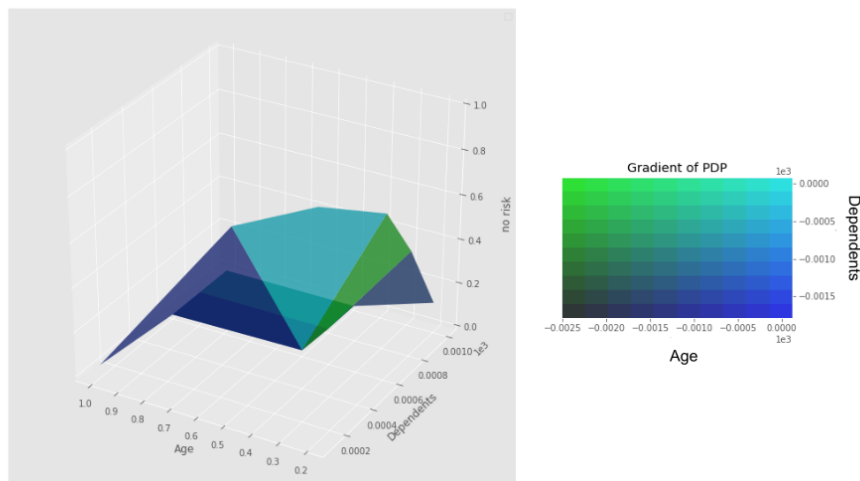


Figure 2.7 Relationship between age, dependents and low-risk output

Fig. 2.7 shows the relationship between AGE, DEPENDENTS and the 'low-risk' class output. The 3D graph on the left shows the features in the axes of the horizontal plane and the output risk class ('low risk') in the vertical axis. The graph on the right explains the colour gradient seen in the PDP. The plot shows how the low-risk taking output changes with different combinations of 'age' and 'dependence'.

This shows how the movement of two features influence the output. In the RegTech tool, the various features whose relationship the regulator wants to observe can be selected and the plots can be created dynamically.

Part 4- Local explanations

'Local' explanations using LIME explain the features that influence a single observation. In the explanation reports, the regulators can randomly select an input condition to understand how the features in that condition affects the output risk class. The report would give the weights of the features influencing the predicted output class (Fig 2.8) and the influence of the input features on all possible output classes (Fig 2.9).

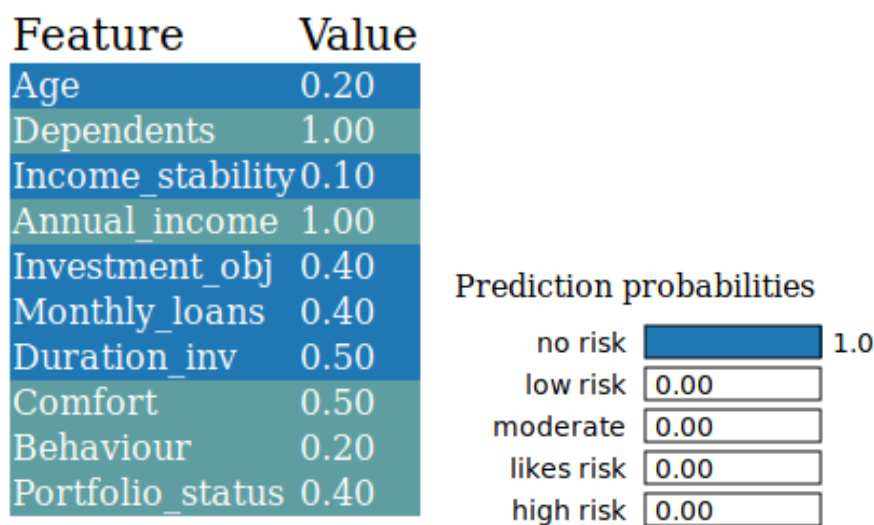


Figure 2. 8 Features influencing predicted output class

Fig. 2.8 shows that for one randomly selected input condition, the table on the left shows the feature values of the input condition and the colour shows the influence it has on the different output classes (shown in the right). It shows that the age, income stability, investment objectives, monthly loans and duration of investment (features in blue) of the user are the primary determining factors that classify the user to the 'no risk' class with a high probability. The features in green (number of dependents, annual income, comfort, behaviour and portfolio status) push the classification towards another output class, however, it's effect is negligible.

In Fig 2.8, the contributions of each feature to and against every class are shown. The highest contributions made by the top features are in the 'no risk' class, all other class contributions are negligible thus the final prediction is 'no risk'.

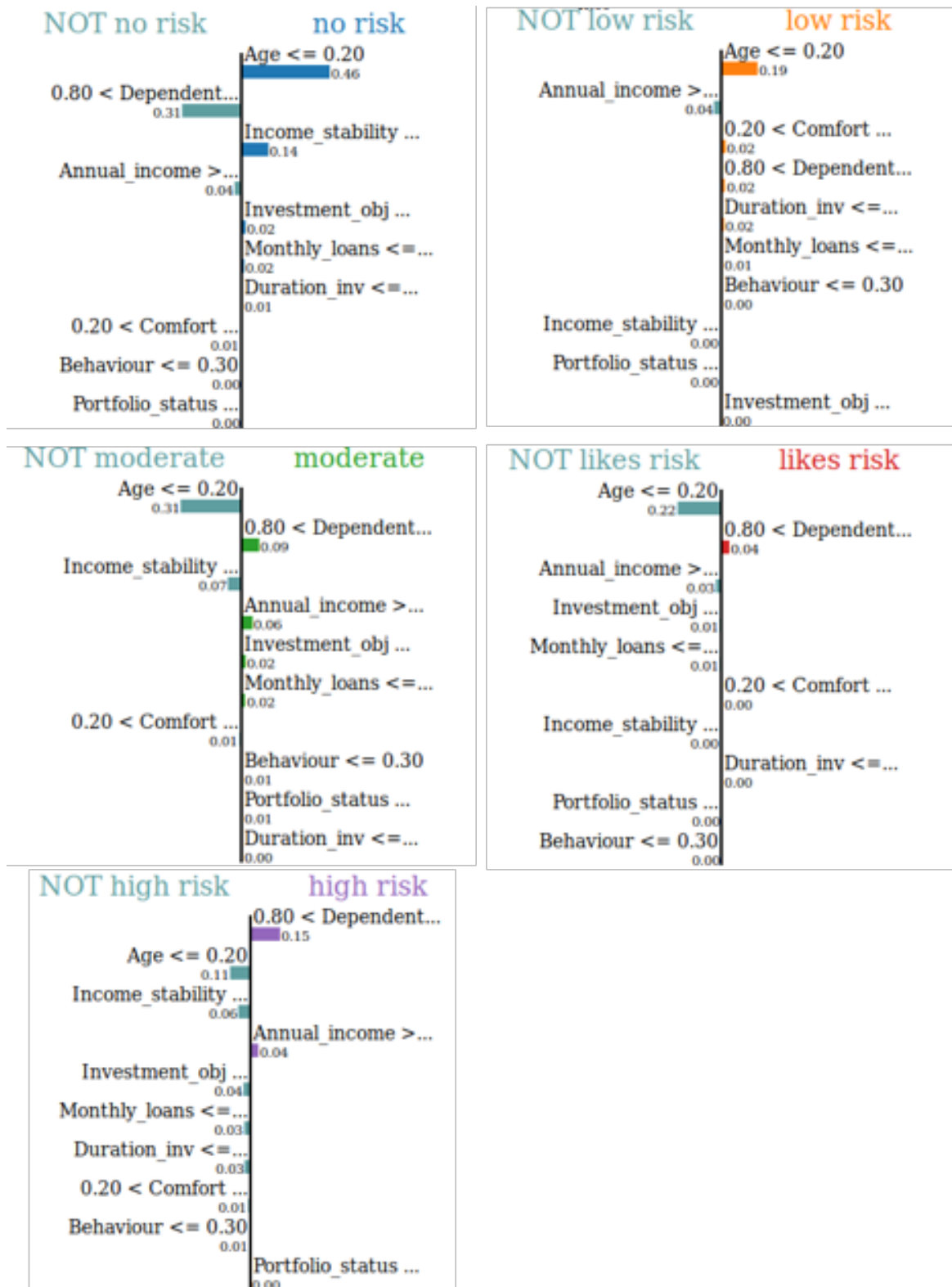


Figure 2.9 Influence of input features on all possible output classes

Fig 2.9 shows class probability for each feature in the observation. For each output class ('no risk' to 'high risk'), each graph shows how the features contribute to the probability of either falling in the class (on the positive axis) or NOT falling in the class (negative axis). In the 'no risk' class, the Age feature strongly pushes it towards 'no risk'.

Fig 2.9 shows how the features in same input condition contribute to the different output classes. In this case, the overall sum of probability lies in the 'no risk' class. The Age feature that matters most has a high probability of belonging to the no risk class.

Using this, regulators can understand a single random observation and understand how the algorithm classifies it to the out class, and hence spot check the algorithms decision making.

Conclusion

In this study, we achieved the following— (i) operationalising explainability in the case of robo advisory risk profiling by creating a RegTech tool that can be used for several algorithms and use cases (ii) describing how this could be used by fintech regulators to audit algorithms and check if they comply with the regulations that they are subject to.

We do this for black box algorithms where firms have to provide a stratified balanced sample of the input-output data, and the regulator uses the RegTech tool to model the data to an equation and generate an audit report with three explanations (consisting of two global and one local explanation method). With this, regulators can understand how each question contributes to the output, how they relate to each other and conduct spot checks. We find that the methods used are able to model the dataset with high degree of accuracy and provide accurate explanations. The methods have been tested using various input conditions to ensure its reliability.

Revisiting the [SEBI rules for automated tools in investment advisory](#), our study has proposed an approach to check if the automated tools comply with the regulations. Using the RegTech tool, we can subject the tool to a comprehensive system audit and inspection. Further, we can provide an explanation for how the tools algorithm works. While an explanation for the algorithm is not mandated, the regulator can use this to check if the robo-advisory tool acts in the best interest of the client without any unintended machine bias.

It is important to note that these explanation methods are replicable for any set on input features, including demographic features (like gender, race), behaviour (such as purchase history or internet activity) or opinions (like political leaning).

Thus, our approach has the potential to enhance the technical capabilities of capital markets regulator without the need for in-house computer science expertise. Considerable work and research would be required to create a comprehensive tool capable of operationalising all regulations.

Discussion and way forward

With algorithms permeating various aspects of public life, they are increasingly being subject to scrutiny and regulations. However, designing and implementing regulations without knowledge of how an algorithmic system works and what its externalities are would prove to be ineffective. To formulate regulations that work, they need to be informed by the technical and operational limitations while also considering the ethical aspects. This is especially true for the case of ADS, where there are glaring problems and yet there is a struggle to enforce concepts like fairness, accountability and transparency. As (Goodman & Flaxman, 2017) point out, the GDPR acknowledges that the few, if any decisions made by algorithms are purely "technical", and the ethical issues posed by them require rare coordination between 'technical and philosophical resources'. Hence, we need dialogue between technologist and regulators and they need to work together to design safeguards by pooling their domain knowledge.

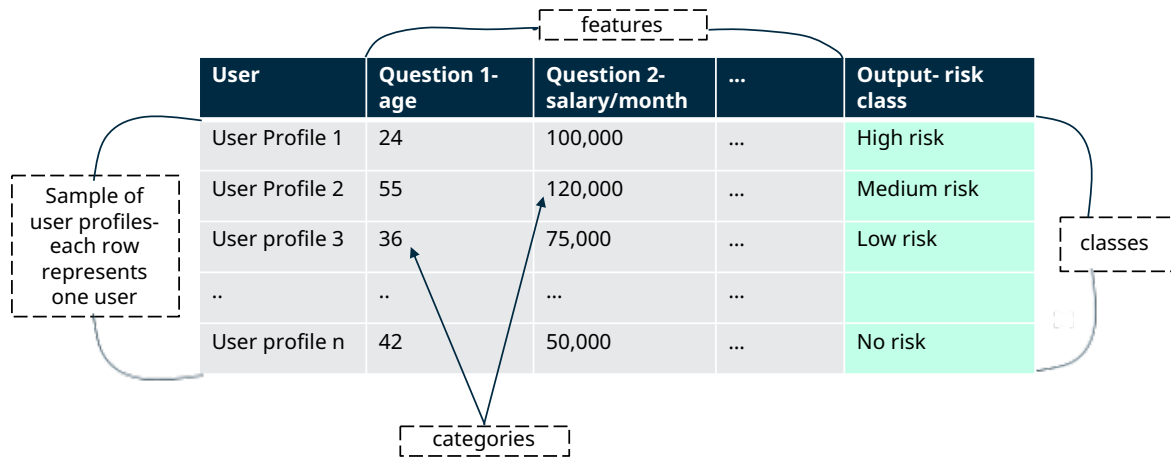
One way to achieve this is by creating regulatory sandboxes. Sandboxes act as test beds where experiments can happen in a controlled environment. They are initiated by regulators for live testing innovations of private firms in an environment that is under the regulator's supervision (Jenik & Lauer, 2017). It can provide a space for dialogue and developing regulatory frameworks for the speed at which technological innovation happens, in a way that "doesn't smother the fintech sector with rules, but also doesn't diminish consumer protection" (BBVA, 2017). This method would help build collaborative regulations and also open up the dialogue of building in explainability by design in ADS early on in the process.

Future work needs to be on the regulatory and technical front. On the regulatory front, we need to work with the regulators to understand the grasp-ability of various explanation methods. Appropriate explanations also need to be extended to the user.

On the technical front, our work can be expanded to include increasingly more complex situations. A standardised and robust documentation process for algorithms also needs to be initiated to maintain accountability and makes it easier to audit the system.

Appendix

Appendix 1- Definitions and key terms



1. Feature- A feature is a measurable property of the object we are trying to analyse. In datasets, features appear as columns¹³.
2. Accuracy- Accuracy gives the percentage of correctly predicted samples out of all the available samples.
Accuracy is not always the right metric to consider in imbalanced class problems; in the risk dataset, class 2 has the most samples greatly outnumbering samples in class 1 and 5. This could mean that even if most samples are incorrectly labelled as belonging to class 2 then the accuracy would still be relatively high giving us an incorrect understanding of the models working. Just considering the accuracy, the most accurate classifier is the decision tree, closely followed by knn and svm who supersede the logistic regression and naive bayes classifiers.
3. Recall- the ability of a model to find all the relevant samples. This gives the number of true positive samples by the sum of true positive and false negative samples. True positive samples are the samples correctly predicted as true by the model and false negatives are data points the model identifies as negative that actually are positive

¹³ <https://www.datarobot.com/wiki/feature/>

(for example points that belong to class 2 that are predicted as not belonging to class 2).

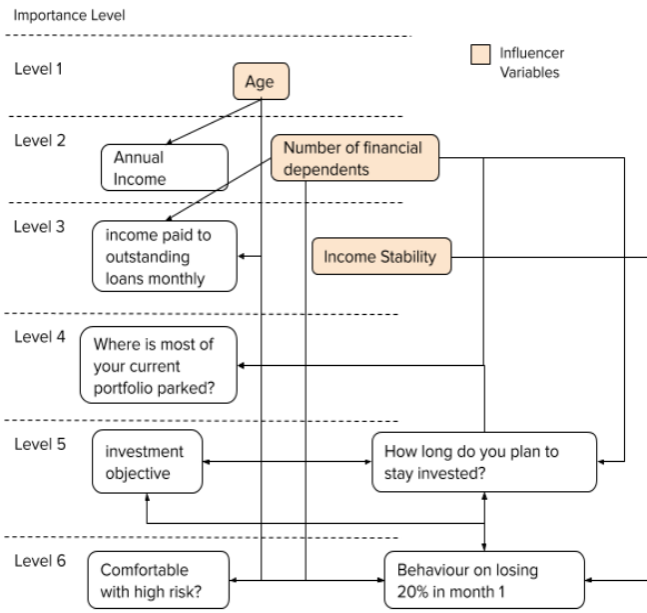
For example, in the performance metrics for logistic regression we find that the performance is thrown off by class moderate/medium -risk takers, this is most probably because the class has too many samples in the training data causing it to overfit (logistic regression is prone to overfitting).

4. Precision- it is defined as the number of true positives divided by the number of true positives plus the number of false positives. False positives are cases the model incorrectly labels as positive that are actually negative, or in our example, individuals the model classifies as class 2 that are not. While recall expresses the ability to find all relevant instances in a dataset, precision expresses the proportion of the data points our model says was relevant actually were relevant.
5. F1 score- Sometimes trying to increase precision can decrease recall and vice versa, an optimal way to combine precision and recall into one metric is by using their harmonic mean also called the F1-Score.

$$F1 = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$$

Appendix 2- Details of sample dataset generation that has been used for this study

We generated a dataset by permuting all possible sequences of the answers for each question (i.e., categories for each feature) asked by prominent robo advisory apps in India. In this case, we used the questions from PayTM money. The flow graph below visualises the importance of the features and the most important variables. table shows the frequently asked questions in robo-advisory apps with corresponding options.

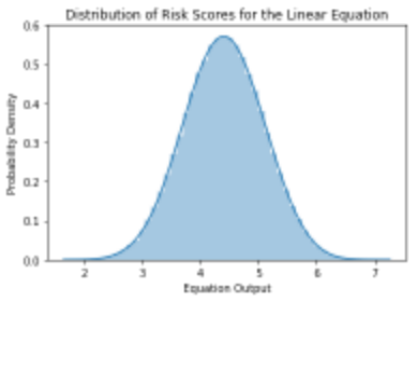
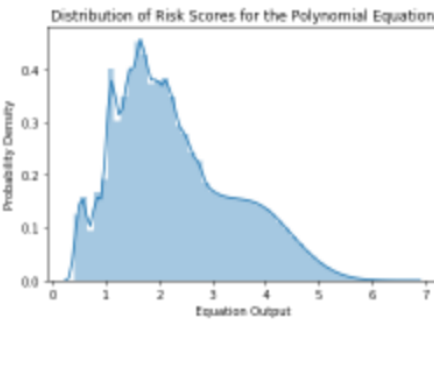

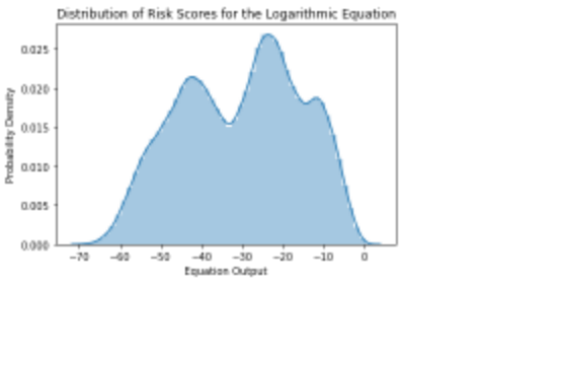
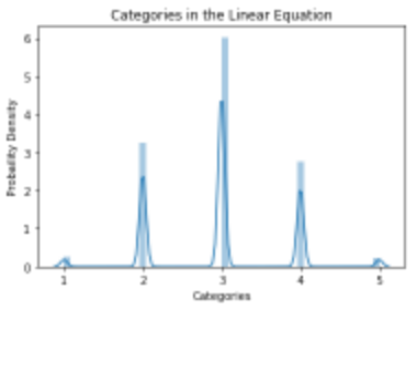
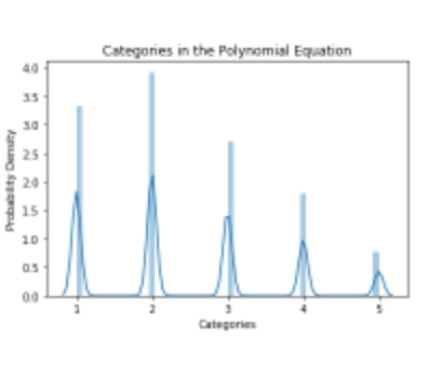
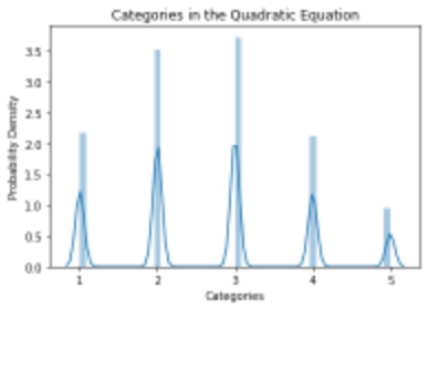
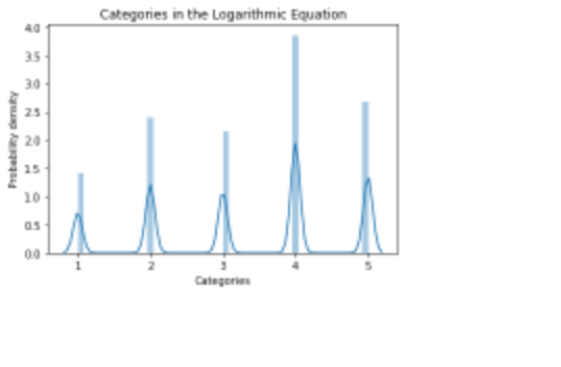


Variable names	Questions	Weight	Option1	Score (option 1)	Option2	Score (option 2)	Option3	Score (option 3)	Option4	Score (option 4)	Option5	Score (option 5)
x(1,1)	What's your age (in years)	1	18-35	1	36-55	0.5	55+	0.2				
x(2,1)	How many people depend on you financially?	0.83	No one	1	Spouse only	0.8	spouse and children	0.6	Parents only	0.6	Spouse, children and parents	0.1
x(2,2)	What's your annual income range?	0.83	Below INR 1 lac	0.2	Between INR 1 Lac - INR 5 Lac	0.4	Between INR 5 lac - 10 Lac	0.6	Between INR 10 Lac - INR 25 Lac	0.8	Above 25 Lac	1
x(3,1)	What % of your monthly income do you pay in outstanding loans, EMI etc?	0.65	None	1	Up to 20% of income	0.8	20-30% income	0.6	30-40% of income	0.4	50% or above of income	0.2
x(3,2)	Please select the stability of your income	0.65	Very low stability	0.1	Low stability	0.3	Moderate Stability	0.6	High Stability	1	Very high stability	1
x(4,1)	Where is most of your current portfolio parked?	0.5	Savings and fixed deposits	0.4	Bonds/debt	0.6	Mutual Funds	0.5	Real Estate or Gold	0.4	Stock Market	0.8
x(5,1)	What's your primary investment objective?	0.8	retirement planning	0.65	Monthly Income	0.6	Tax Saving	0.4	Capital Preservation	0.5	Wealth Creation	1

x(5,2)	How long do you plan to stay invested?	0.8	Less than 1 year	0.5	1 to 3 years	0.8	3 to 5 years	0.65	5 to 10 years	0.6	more than 10 years	0.7
x(6,1)	To achieve high returns, you are comfortable with high risk investments	0.7	Strongly agree	1	Agree	0.9	Neutral	0.5	Disagree	0.2	Strongly disagree	0.1
x(6,2)	If you lose 20% of your invested value one month after investment, you will	0.65	Sell and preserve cash	0.2	Sell and move cash to fixed deposits or liquid fund	0.3	Wait till market recovers and then sell	0.5	Keep investments as they are	0.8	Invest more	1

Table above - Frequently asked questions in robo-advisory apps with corresponding options. The weights to the questions (features) and scores given to the options (categories) were set at our discretion in order to generate the dataset. The values are for representation and the method would work for any set values.

Types of Equations	Linear Equation under Independent Variable assumption	Equation with interaction effects	Quadratic Equation	Logarithmic Equation
Equations	$w_{11} \cdot \text{Age} + w_{21} \cdot \text{Dependents} + w_{22} \cdot \text{Annual_Income} + w_{31} \cdot \text{Monthly_loans} + w_{32} \cdot \text{Income_stability} + w_{41} \cdot \text{Portfolio_status} + w_{51} \cdot \text{Investment_obj} + w_{52} \cdot \text{Duration_inv} + w_{61} \cdot \text{Comfort} + w_{62} \cdot \text{Behaviour} = \text{output}$	$w_{11} \cdot \text{Age} + w_{21} \cdot \text{Dependents} + w_{22} \cdot [k] \cdot \text{Age} + w_{31} \cdot x_{31}[l] \cdot \text{Age} \cdot \text{Dependents} + w_{32} \cdot x_{32}[m] + w_{41} \cdot x_{41}[n] \cdot \text{Age} \cdot \text{Dependents} + x_{52}[o] + w_{51} \cdot x_{51}[o] \cdot \text{Age} \cdot \text{Dependents} + x_{52}[p] + w_{52} \cdot x_{52}[p] \cdot x_{51}[o] \cdot \text{Age} \cdot \text{Dependents} + w_{61} \cdot x_{61}[q] \cdot \text{Age} \cdot \text{Dependents} + w_{62} \cdot x_{62}[r] \cdot \text{Age} \cdot \text{Dependents} \cdot x_{32}[m] \cdot x_{52}[p]$	$w_{11}(\text{Age}^{** 3}) + w_{21} \cdot \text{Age} \cdot (\text{Dependents}^{* 2}) + w_{22} \cdot \text{Age} \cdot \text{Annual_Income} + w_{31}(\text{Monthly_loans}^{** 2}) + w_{32}(\text{Income_stability}^{* 3}) + w_{41} \cdot \text{Dependents} \cdot \text{Portfolio_status} + w_{51}(\text{Investment_obj}^{** 2}) \cdot \text{Monthly_loans} + w_{52} \cdot \text{Duration_inv} \cdot \text{Dependents} + w_{61} \cdot \text{Monthly_loans} \cdot \text{Comfort} + w_{62} \cdot \text{Behaviour} \cdot \text{Dependents}$	$w_{11} \cdot 3 \cdot \text{math.log}(\text{Age}, 3) + w_{21} \cdot 2 \cdot \text{math.log}(\text{Age} \cdot \text{Dependents}, 2) + w_{22} \cdot 3 \cdot \text{math.log}(\text{Age} \cdot \text{Annual_Income}, 2) + w_{31} \cdot 3 \cdot \text{math.log}(\text{Age} \cdot \text{Monthly_loans}, 2) + w_{32} \cdot 3 \cdot \text{math.log}(\text{Age} \cdot \text{Income_stability}, 2) + w_{41} \cdot \text{Portfolio_status} + w_{51} \cdot 3 \cdot \text{math.log}(\text{Age} \cdot \text{Investment_obj}, 2) + w_{52} \cdot \text{Duration_inv} \cdot \text{Behaviour} + w_{61} \cdot 2 \cdot \text{math.log}(\text{Comfort} \cdot \text{Age}, 2) + w_{62} \cdot \text{Behaviour} \cdot \text{Age} = \text{output}$
Range of outputs [min, max]	[1.764, 7.15]	[0.394, 6.74]	[0.18, 7.15]	[-69.70, 1.69]

<p>Distribution of risk scores</p>				
<p>Boundaries</p>	<ul style="list-style-type: none"> ▪ No risk: less than 3 ▪ Low risk: 3 to 4 ▪ Moderate risk: 4.1 to 4.9 ▪ Likes risk: 5 to 5.8 ▪ High risk: more than 5.8 	<ul style="list-style-type: none"> ▪ No risk: less than 1.5 ▪ Low risk: 1.6 to 2.3 ▪ Moderate risk: 2.4 to 3.3 ▪ Likes risk: 3.4 to 4.3 ▪ High risk: more than 4.3 	<ul style="list-style-type: none"> ▪ No risk: less than 1.5 ▪ Low risk: 1.6 to 2.3 ▪ Moderate risk: 2.4 to 3.3 ▪ Likes risk: 3.4 to 4.3 ▪ High risk: more than 4.3 	<ul style="list-style-type: none"> ▪ No risk: less than -50 ▪ Low risk: -49 to -40 ▪ Moderate risk: -39 to -30 ▪ Likes risk: -30 to -17 ▪ High risk: more than -17
<p>After boundary class category distribution</p>				
<p>Total number of observations</p>	<ul style="list-style-type: none"> ▪ No risk : 1 : 60,923 ▪ Low risk : 2 : 8,17,511 ▪ Moderate risk : 3 : 15,15,986 	<ul style="list-style-type: none"> ▪ No risk : 1 : 9,96,032 ▪ Low risk : 2 : 11,76,069 ▪ Moderate risk : 3 : 8,08,223 	<ul style="list-style-type: none"> ▪ No risk : 1 : 6,53,408 ▪ Low risk : 2 : 10,55,754 ▪ Moderate risk : 3 : 11,18,259 	<ul style="list-style-type: none"> ▪ No risk : 1 : 4,22,859 ▪ Low risk : 2 : 7,17,505 ▪ Moderate risk : 3 : 4,22,859 ▪ Likes risk : 4 : 11,56,591

in each category	<ul style="list-style-type: none"> ▪ Likes risk: 4 : 6,90,604 ▪ High risk: 5 : 60,701 	<ul style="list-style-type: none"> ▪ Likes risk: 4 : 5,36,121 ▪ High risk: 5 : 2,33,555 	<ul style="list-style-type: none"> ▪ Likes risk: 4 : 6,37,694 ▪ High risk: 5 : 2,84,885 	<ul style="list-style-type: none"> ▪ High risk: 5 : 8,04,616
Final data sample chosen for models (stratified/solving the imbalanced class problem)	9,43,718 rows of data. <ul style="list-style-type: none"> ▪ No risk : 1 : 18,277 ▪ Low risk : 2 : 2,45,254 ▪ Moderate risk : 3 : 4,54,796 ▪ Likes risk : 4 : 2,07,181 ▪ High risk : 5 : 18,210 	11,25,000 rows of data. <ul style="list-style-type: none"> ▪ No risk : 1 : 2,98,810 ▪ Low risk : 2 : 3,52,821 ▪ Moderate risk : 2,42,467 ▪ Likes risk : 4 : 1,60,836 ▪ High risk : 5 : 70,066 	11,25,000 rows of data. <ul style="list-style-type: none"> ▪ No risk : 1 : 196022 ▪ Low risk : 2 : 316726 ▪ Moderate risk : 335478 ▪ Likes risk : 4 : 191308 ▪ High risk : 5 : 85466 	11,25,000 rows of data. <ul style="list-style-type: none"> ▪ No risk : 1 : 1,26,858 ▪ Low risk : 2 : 2,15,251 ▪ Moderate risk : 3 : 1,94,529 ▪ Likes risk : 4 : 3,46,977 ▪ High risk : 5 : 2,41,385
Correlations of variables with final category column	<ul style="list-style-type: none"> ▪ Age 0.453045 ▪ Dependents 0.374792 ▪ Annual_income 0.404084 ▪ Monthly_loans 0.232607 ▪ Income_stability 0.302768 ▪ Portfolio_status 0.103018 ▪ Investment_obj 0.222040 ▪ Duration_inv 0.109305 ▪ Comfort 0.345110 ▪ Behaviour 0.283973 ▪ output 0.931334 	<ul style="list-style-type: none"> ▪ Age 0.742770 ▪ Dependents 0.516709 ▪ Annual_income 0.109562 ▪ Monthly_loans 0.053320 ▪ Income_stability 0.214954 ▪ Portfolio_status 0.014864 ▪ Investment_obj 0.063204 ▪ Duration_inv 0.036067 ▪ Comfort 0.074225 ▪ Behaviour 0.018715 ▪ output 0.966746 ▪ categories 1.000000 	<ul style="list-style-type: none"> ▪ Age 0.640698 ▪ Dependents 0.485437 ▪ Annual_income 0.109455 ▪ Monthly_loans 0.371431 ▪ Income_stability 0.213423 ▪ Portfolio_status 0.040315 ▪ Investment_obj 0.118964 ▪ Duration_inv 0.043781 ▪ Comfort 0.131617 ▪ Behaviour 0.114555 ▪ output 0.963646 ▪ categories 1.000000 	<ul style="list-style-type: none"> ▪ Age 0.849599 ▪ Dependents 0.119673 ▪ Annual_income 0.110896 ▪ Monthly_loans 0.088254 ▪ Income_stability 0.318025 ▪ Portfolio_status 0.003605 ▪ Investment_obj 0.057762 ▪ Duration_inv 0.004719 ▪ Comfort 0.094796 ▪ Behaviour 0.015175 ▪ output 0.971413 ▪ categories 1.000000

	<ul style="list-style-type: none">categories 1.000000			
--	---	--	--	--

Table above - Details of the sample input-output data generated for the study using four equations.

Appendix 3- Explaining the machine learning models

Logistic Regression

Logistic Regression is a commonly used statistical method for analysing and predicting data with one or more independent variables and one binary dependent variable; for example, spam or not spam email classifiers, benign or malignant tumour detection. A logistic regression classifier tries to fit data according to a linear hypothesis function such as $Y = W(i)x(i) + B$ (Similar to a line equation) where Y is the dependent variable, X represents independent variables from 1 to n , B gives an error bias (negligible) and W is the weight assigned to each variable. W is an important value as it tells us the individual contributions of variables in determining Y , our target.

The independent variable is always binary, in our case there will be five logistic regression classifiers with their independent variables as 1 (Low Risk) or Not 1 (Not Low Risk), 2 or Not 2 and so forth till case 5 (High Risk). This format of multiclass classification is called 'one versus rest', the input sample is passed through all the classifiers and probability of the sample belonging to classes 1 to 5 is calculated and the highest probability class wins.

The interpretation of weights in logistic regression is dependent on the probability of class classification, the weighted sum is transformed by the logistic function to a probability. Therefore, the interpretation equation is:

$$\log\left(\frac{P(y = 1)}{1 - P(y = 1)}\right) = \log\left(\frac{P(y = 1)}{P(y = 0)}\right) = \beta_0 + \beta_1x_1 + \dots + \beta_px_p$$

The log function calculates the odds of an event occurring.

$$\frac{P(y = 1)}{1 - P(y = 1)} = odds = \exp(\beta_0 + \beta_1x_1 + \dots + \beta_px_p)$$

Logistic regression is used over linear regression as completely linear model do not give output probabilities because it treats the classes as numbers (0 and 1) and fits the best hyperplane (for a single feature, it is a line) that minimises the distances between the points and the hyperplane. In other words, it simply interpolates between the points, and we cannot interpret it as probabilities. A linear model also extrapolates and gives us values below zero and above one. Logistic regression is also widely used, interpretable and fits our use case relatively well.

Support Vector Machine (SVM) Classifier

A support vector machine finds an equation of a hyper-plane that separates two or more classes in a multidimensional space; for example, if we consider a two-dimensional space, this "hyperplane" will become a line dividing the plane on which the data lies into two separate classes. If the data is not linearly separable i.e. there is no clear line separating the classes (This happens in many cases; imagine two classes in the data forming concentric circles) then data can be transformed onto a different plane (say we view the concentric circles from z axis) it becomes a linearly separable problem again (imagine the points in the circle having different depth). After separating it we can transform it back to the original plane: this is done using a kernel function in SVM.

Support vector machines have become wildly popular due to their robust efficiency and high accuracy despite requiring very few samples to train. They have disadvantages especially when it comes to time and space complexity but the SVM algorithm along with its variations are being used commercially in face detection and protein fold predictions.

SVM for multiclass classification trains $n*(n-1)/2$ classifiers, where n is the number of classes in the problem. Therefore, for our problem there will be 10 different classifiers each will choose permutations of classes as the binary dependent variable (Y) i.e., 1 or 2, 2 or 3, 1 or 4 and all others. During this, each classifier predicts one class instead of probabilities for each.

Interpreting the above is quite difficult, the benefit of a linear model was that the weights / parameters of the model could be interpreted as the importance of the features. But if the model is non-linear, it would not work. Once we engineer a high or infinite dimensional feature set, the weights of the model implicitly correspond to the high dimensional space which isn't useful in aiding our understanding of SVM's. What we can do is fit a logistic regression model which estimates the **probability** of label Y being 1, given the original features. We use maximum likelihood estimation to fit the parameters of the logistic regression model, the technique is called Platt Scaling.

For our use case we use a kernel with interaction effects for learning hyperplane boundaries as our original equation used to generate data is correlated in a equation with interaction effects, but this adds some more complexity to the algorithm. The kernel with interaction effects can be written as $K(x, x_i) = 1 + (x x_i) / d$; where x is the input vector and x_i represents support vectors (hyperplane equations).

Decision Tree classifier

Decision trees belong to the family of tree-based learning algorithms, they are widely used for supervised classification as they create precise, well defined and hierarchical decision boundaries for categorical and continuous data. This differs from classifiers that use a single separation boundary (or line) such as logistic regression by iteratively splitting the data into subparts by identifying multiple divisive boundaries.

The conditions that make these divisions try to ensure an absence of impurities in the populations contained by them; for example, a condition that decision tree will make to describes a 'banana' could be in the sequence type=" fruit", colour = "yellow", shape = "crescent", spots = "true" this leaves no place for uncertainty or impurity. The algorithm stops when all classes are pure or there are no features left to divide upon.

Unfortunately, such sharp dividing conditions are not always possible or may exceed certain time and space limitations in real life. Therefore, when a clear separation of classes is not possible then we can have a stopping condition that tolerates some impurity (For example gini impurity measures quality of such splits by calculating the probability of an incorrect classification of a randomly picked datapoint).

The impurity itself can be calculated using a measure of randomness, **entropy**: $H = -p(x)\log(p(x))$ or $-p\log(p) - q\log(q)$ where p = probability of success and q = prob of failure

Ideally H should be as small as possible.

For a dataset like ours with multiple features, deciding the splitting feature i.e. most important dividing condition at each step is a complex task, this feature should reduce the impurity through the split or one with gives the most information gain. **Information gain** at each node is calculated by the lowest entropy generated nodes by the split. Starting from the root node, you go to the next nodes and the edges tell you which subsets you are looking at. Once you reach a leaf node, the node tells you the predicted outcome. All the edges are connected by 'AND'. For example: If feature x is [smaller/bigger] than threshold c AND etc... then the predicted outcome is the mean value of y of the instances in that node.

Individual decisions made by the tree can also be explained by going down a particular path based on the input given. Decision trees can be used to explain the dataset by themselves.

Naïve Bayes

Naive Bayes classifiers are a family of classifiers that work on predicting future outcomes using conditional probability, given a history of behaviour. For example, given a year long history of weather forecasts with features such as humidity, rainfall, and temperature, a classifier from the naive Bayes family can be trained and used to predict future weather conditions. Due to its simplicity, it has found a place in many real-world systems such as credit scoring systems,

weather prediction and many others. Given its popularity, we have used it to model our dataset.

The Bayes algorithm works under a “naive” assumption that all the features are independent in nature, in our case that means the naive Bayes classifier is going to assume that our variables such as age, income are uncorrelated so finding probabilities can be thought of as a simple counting calculation. This implies that the classifier won't be a right fit for our case as we know that the data was generated using many correlations (such as age will affect an individual's income, behaviour etc..).

If the naive Bayes classifier wants to calculate the probability of observing features f_1 to f_n , given a class c (In our case c here, represents the risk class and f values represent all our question-answer scores), then

$$p(f_1, f_2, \dots, f_n | c) = \prod_{i=1}^n p(f_i | c)$$

This means that when Naive Bayes is used to classify a new example, the posterior probability is much simpler to work with:

$$p(c | f_1, f_2, \dots, f_n) \propto p(c) p(f_1 | c) \dots p(f_n | c)$$

But we have left $p(f_n | c)$ undefined i.e. the occurrence of a certain feature given a class which means we haven't taken the distribution of the features into account yet. Therefore, for our case we have used a **gaussian naive Bayes** classifier that simply assumes $p(f_n | c)$ is a gaussian normal distribution, this works well for our data which is a normal distribution.

Then the formula for our low-risk class used by the classifier will be something like:

$P(\text{low-risk} / \text{Age, Income, Dependents...}) = P(\text{low-risk} / \text{Age-category}) * P(\text{low-risk} / \text{Income-category}) \text{ etc} / P(\text{Age}) * P(\text{income}) \text{ etc}$. This will be calculated for all risk categories and the class with the highest probability is given as the final prediction.

Naive Bayes is an interpretable model because of the independence assumption. It can be interpreted on the modular level. The contribution made by each feature towards a specific class prediction is clear, since we can interpret the conditional probability.

K-Nearest Neighbours (KNN)

Neighbours-based classification is a type of *instance-based learning* or *non-generalizing learning*: it does not try to construct a general internal model, but simply stores instances of the training data. In KNN, a data point is classified by a majority vote of its neighbours. The input is assigned the class most common among its 'k' nearest neighbours, where 'k' is a small positive integer, the value of 'k' is chosen depending on the data. KNN is very useful in applications that require searching for similar items; such as recommender systems, bio-surveillance software, document retrieval systems such as concept search which is used in many e-Discovery software packages.

These neighbours are decided using brute force techniques that calculate distance from the data point of interest to all the other data points in the dataset, by using formulae like Euclidean distance. This means that the time and space complexity of this operation is very high; for n samples in d dimensions the time complexity will be $O(d*n*n)$ which makes this algorithm relatively slow to run on large datasets.

Since KNN is an instance-based algorithm there is no learned model, there are no parameters to learn, so there is no interpretability on a modular level. There is a lack of global model interpretability because the model is inherently local and there are no global weights or structures explicitly learned. To explain a prediction at a local level, we can always retrieve the k neighbours that were used for the prediction. This is useful for our dataset as there will be thousands of neighbouring data points but presenting those 'k' nearest points could be a very useful explanation for each category.

References

- Narayanan, A. (2016, June 27). *Investor Business Daily*. Retrieved October 2019, from <https://www.investors.com/etfs-and-funds/etfs/fund-industry-wakens-from-slumber-to-take-on-digital-advice-upstarts/>
- Carey, T. (2019, September 24). *Investopedia*. Retrieved October 2019, from <https://www.investopedia.com/robo-advisors-2019-where-have-all-the-assets-gone-4767826>
- Goodman, B., & Flaxman, S. (2017). *European Union regulations on algorithmic decision-making and a "right to explanation"*. Retrieved October 2019, from https://ora.ox.ac.uk/catalog/uuid:593169ee-0457-4051-9337-e007064cf67c/download_file?file_format=pdf&safe_filename=euregs.pdf&type_of_work=Journal+article
- EU GDPR. (2016). *EU GDPR Chapter 3*. Retrieved October 2019, from <https://gdpr.eu/article-22-automated-individual-decision-making/>
- Wachter, S., Mittelstadt, B., & Floridi, L. (2016, December 28). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law 2017*.
- Shneiderman, B. (2017, May 30). *Algorithmic Accountability*. The Alan Turing Institute.
- Citron, D. K., & Pasquale, F. A. (2014). The Scored Society: Due Process for Automated Predictions. *Washington Law Review*, 89, 1-34.
- Kapur, D., & Khosla, M. (2019). *Regulation in India: Design, Capacity, Performance*. Hart Studies in Comparative Public Law.
- Padmanabhan, A., & Rastogi, A. (2019). Big Data. In D. Kapur, & M. Khosla, *Regulation in India: Design, Capacity, Performance* (pp. 251-278). Hart Studies in Comparative Public Law.
- Ciocca, P., & Biancotti, C. (2018, October 23). Data superpowers in the age of AI: A research agenda. *VOX CEPR Portal*.

- Thelisson, E., Padh, K., & Celis, E. L. (2017, July 15). Regulatory Mechanisms and Algorithms towards Trust in AI/ML.
- Cowls, J., King, T., Taddeo, M., & Floridi, L. (2019, May 15). Designing AI for Social Good: Seven Essential Factors. *SSRN*
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3388669.
- Rudin, C. (2019, May 13). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 206-215.
- IOSCO. (2014, July). *Report on the IOSCO Social Media and Automation of Advice Tools Surveys*. Retrieved September 2019, from
<https://www.iosco.org/library/pubdocs/pdf/IOSCOPD445.pdf>
- Kaya, O. (2017, August 10). *Robo-advice – a true innovation in asset management*. Retrieved September 2019, from
https://www.dbresearch.com/PROD/RPS_EN-PROD/PROD0000000000449125/Robo-advice_%E2%80%93_a_true_innovation_in_asset_managemen.pdf
- Abraham, F., Schmukler, S. L., & Tessada, J. (2019, February). *Robo-Advisors: Investing through Machines*. Retrieved October 2019, from
<http://documents.worldbank.org/curated/en/275041551196836758/text/Robo-Advisors-Investing-through-Machines.txt>
- FINRA. (2016, March). *Report on Digital Investment Advice*. Retrieved September 2019, from FINANCIAL INDUSTRY REGULATORY AUTHORITY:
<https://www.finra.org/sites/default/files/digital-investment-advice-report.pdf>
- Dastin, J. (2018, October 10). *Amazon scraps secret AI recruiting tool that showed bias against women*. (Reuters) Retrieved September 2019, from
<https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MKo8G>

- ProPublica. (2016, May 23). *Machine Bias There's software used across the country to predict future criminals. And it's biased against blacks*. Retrieved September 2019, from <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- Datta, A., Tschantz, M. C., & Datta, A. (2015, February 18). Automated Experiments on Ad Privacy Settings. *Proceedings on Privacy Enhancing Technologies* , 92-112.
- Costello, M., Hawdon, J., Ratliff, T., & Grantham, T. (2016, May). Who views online extremism? individual attributes leading to exposure. *Computers in Human Behavior*.
- Baer, D. (2019, November). *The 'Filter Bubble' Explains Why Trump Won and You Didn't See It Coming*. Retrieved October 2019, from The Cut: <https://www.thecut.com/2016/11/how-facebook-and-the-filter-bubble-pushed-trump-to-victory.html>
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019, October 25). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*.
- Kari, P. (2019, October 25). *Healthcare algorithm used across America has dramatic racial biases*. Retrieved October 2019, from Guardian: <https://www.theguardian.com/society/2019/oct/25/healthcare-algorithm-racial-biases-optum>
- Castelluccia , C., & Le Métayer, D. (March 2019). *Understanding algorithmic decision-making: Opportunities and challenges*. Study, European Parliamentary Research Services, Panel for the Future of Science and Technology.
- Maurell, v. d. (2019). *Embracing Robo Advisory looks promising or the longitivity of Financial Advisors*. Global Financial Markets Institute, New York.
- Wired.com. (2019, November 19). *The apple card didn't see gender and that's the problem*. Retrieved December 2019, from

<https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/>

Hall, P., & Gill, N. (2018). *An Introduction to Machine Learning Interpretability*. (N. Tache, Ed.) O'Reilly.

BBVA. (2017, November 18). *What is a regulatory sandbox?* Retrieved December 2019, from <https://www.bbva.com/en/what-is-regulatory-sandbox/>

Jenik, I., & Lauer, K. (2017). *Regulatory Sandboxes and Financial Inclusion*. CAGP. <https://www.cgap.org/sites/default/files/Working-Paper-Regulatory-Sandboxes-Oct-2017.pdf>.

Laboure, M., & Braunstein, J. (2017, November 11). *Democratising finance: The digital wealth management revolution*. Retrieved October 2019, from VOX CEPR Policy Portal: <https://voxeu.org/article/digital-wealth-management-revolution>

SEBI. (2016, December 8). *SEBI (Investment Advisers) Regulations 2013 [Last amended on December 08, 2016]*. Retrieved 2019 August, from [sebi.gov.in: https://www.sebi.gov.in/legal/regulations/jan-2013/sebi-investment-advisers-regulations-2013-last-amended-on-december-08-2016-_34619.html](https://www.sebi.gov.in/legal/regulations/jan-2013/sebi-investment-advisers-regulations-2013-last-amended-on-december-08-2016-_34619.html)

SEBI. (2016, October 26). *Consultation Paper on Amendments/Clarifications to the SEBI (Investment Advisers) Regulations, 2013*. Retrieved August 2019, from [sebi.gov.in: https://www.sebi.gov.in/sebi_data/attachdocs/1475839876350.pdf](https://www.sebi.gov.in/sebi_data/attachdocs/1475839876350.pdf)

Tutt, A. (2016, March 15). An FDA for Algorithms. *Administrative Law Review*.

Mulgan, G. (2016, February). A machine intelligence commission for the UK: how to grow informed public trust and maximise the positive impact of smart machines. *Nesta*.

Sample, I. (2017, January 27). *This article is more than 3 years old AI watchdog needed to regulate automated decision-making, say experts*. Retrieved January 2020, from The Guardian:

<https://www.theguardian.com/technology/2017/jan/27/ai-artificial-intelligence-watchdog-needed-to-prevent-discriminatory-automated-decisions>

- Andrews, L. (2017). Algorithms, governance and regulation: beyond 'the necessary hashtags'. In LSE, *Algorithmic Regulation* (pp. 7-12). London.
- New, J., & Castro, D. (2018). *How Policymakers Can Foster Algorithmic Accountability*. Center for Data Innovation.
- Chazot, C. (2015, October). (R. E. INSTITUTE OF INTERNATIONAL FINANCE, Interviewer)
- Arner, D., Barberis, J., & Buckley, R. (2016). *FinTech, RegTech and the Reconceptualization of Financial Regulation*.
- Diakopoulos, N., Friedler, S., Arenas, M., Barocos, S., Hale, M., Howe, B., . . . Zevenbergen, B. (n.d.). *Principles for Accountable Algorithms*. Retrieved December 2019, from FAT ML:
<https://www.fatml.org/resources/principles-for-accountable-algorithms>
- Bracke, P., Datta, A., Jung, C., & Sen, S. (2019). *Machine learning explainability in finance: an application to default risk analysis*.
<https://www.bankofengland.co.uk/-/media/boe/files/working-paper/2019/machine-learning-explainability-in-finance-an-application-to-default-risk-analysis.pdf>, Bank of England.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August 9). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *arxiv*.

Algorithmic Explainability in Practice: Evaluating the effectiveness of explanations in the context of robo advisory apps¹⁴

Abstract

Robo financial advisors (RAs) are complex algorithmic decision-making systems, which gained prominence with their claim to “democratise” finance. Due to their low operating costs and multi-tasking abilities, RAs have the potential for mass adoption. At the same time, it has been seen that the lack of transparency and explanations for these automated decisions leads to a trust deficit for users. The primary aim of this paper is to analyse the effectiveness of user-centric explanations in conveying the decision-making logic of complex algorithmic systems, in this case, for RAs. We surveyed several categories of Ras and used a questionnaire to build a dataset containing all potential client characteristics. Our approach tests these algorithms using techniques from explainable AI to bridge this trust deficit by conducting a user study with 105 human subjects on a custom-built RA. The quantitative aspects of our study determine the efficacy and usability of explanations and the qualitative aspects measure the effect of explanations on users and system usability, also highlighting the need for such an explanation strategy. Our study finds that users show high comprehension and a positive response towards explanations regardless of their technical nature. Trust and confidence of users in the system is positively correlated with comprehension as well as the presence of an explanation. There is a notable reduction in comprehension and trust between transparent white and opaque black box explanations of algorithms. This study is designed to aid policymakers and regulators in order to understand user needs which are crucial to design better policies around algorithmic explainability for RAs.

Keywords: algorithmic decision-making systems (ADS), artificial intelligence, algorithmic regulation, algorithmic explainability and transparency, robo

¹⁴ An edited version of this paper has been published under the title, “User-Centric Explainability in Fintech Applications”.

https://doi.org/10.1007/978-3-030-78642-7_64

financial advisory apps, fintech, explainable AI (XAI), human computer interaction, UX Research, UX Design, explainability by design

Introduction

The financial services sector is one of the torchbearers for applications driven by artificial intelligence (AI). Artificial intelligence is used in this sector for high-quality customer engagement through personalisation, virtual customer service and chatbots. AI has helped improve processes in back-office operations through intelligent automation, and data analysis from bank and social media records have allowed for new ways to measure credit-worthiness. On the security front, proactive monitoring and better prevention of fraud, money laundering and other malpractices have become possible because of AI. However, the application of AI in wealth management, through the routes of robo-advisory, algorithmic trading and automated transactions, have raised pressing questions with respect to the potential negative effects of AI on individuals. This paper's objective is to bridge this gap by testing the effectiveness of user-centric explanations in conveying the decision-making logic of complex algorithmic systems, for robo-advisory applications. Our approach tests these algorithms using techniques from explainable AI to bridge this trust deficit by conducting a user study with 105 human subjects on custom-built robo-advisors.

This is the final paper in a series of three papers on the subject of AI in the financial services sector. While the first paper in this series examined the trade-off between algorithmic performance and explainability, the second paper operationalised the concept of explainability, by building a tool to generate explanations that satisfy regulatory and user requirements. The third paper now dives deeper into the nature of algorithmic explainability in practice, by evaluating the effectiveness of explanations in the context of robo advisory applications.

This paper begins by describing the need and demand for the robo-advisory use case through Section I, followed by the challenging trust deficit hindering its mass adoption. Through Section II, the literature review, we define explainable AI (XAI) and its key evolutionary developments. The review concludes by explaining user information requirements from algorithmic systems that drive our user research, gathered through relevant literature and research on XAI. Section III defines the research questions and objectives addressed through our explanation strategy and user study. Section IV describes the methodology followed by the study divided into six stages. First, we define the scope of our study which analyses the results of a real task solved by real humans. Second, we describe the major components of the robo advisory system built for the experiment; user risk profiling and mutual fund recommendation. Third, we analyse and present the questionnaire used to classify user risk behaviour. The fourth part delves into the explanation strategies used for white and black box systems. Fifth, the objectives and procedure followed by the user study is explained. Sixth, we categorise each question asked in the user survey according to their purpose — whether to assess user comprehension, to satisfy user information requirements or to gather user thoughts and opinions.

Section V analyses the results of the user survey. The results are divided into seven parts. This section details the user study participants and their demographics, analysis of user comprehension and opinions to the survey. Based on this we assess the usability of explanations and the system as a whole. We also analyse the results based on different demographic groups of participants. Finally, in Section VI, we conclude the paper by summarising our major findings and future research objectives. One major finding suggests that users are well-equipped to understand explanations of a complex algorithms, particularly explanations that provide a personalised but partial overview of how the system uses features. Additionally, these explanations are found to be positively correlated with user trust and consequently usage of the system. Therefore, explanations of decision logic would benefit users as well as the developers of these systems. Future work might include using insights gained here to design a

more generalisable strategy of explanations, complemented by cross-domain (legal, ethical, and policy) support.

Emergence of Robo Financial Advisors

Robo-advisory applications are online investment advisory algorithms that are automated and designed to recommend “the best plans for trading, investment, portfolio rebalancing, or tax saving, for each individual as per their requirements and preferences” (Krishnan, et al., 2020). The process generally involves the client responding to a questionnaire and subsequently being categorised on a spectrum of risk ranging from low to high (Krishnan, et al., 2020). Robo-advisories then use the information provided to offer financial advice for a range of goals such as retirement, education or an exigency fund (Bank, 2019). Robo-advisers are considered to be cheaper and more accessible when compared to human advisors, and hold the potential to democratise financial services by providing financial advice to sections of the population which are currently outside the formal banking system (Krishnan, et al., 2020). This would help to arrest the consumer-producer gap in terms of the spatial and temporal dimensions that is prevalent among the current intermediaries in the financial system (banks, agents etc). The ability of these applications to devise region and culture-specific investment strategies allows financial products to be made adaptable to local conditions.

Robo-advisory applications have found their footing in the Indian financial services sector as well. Indian fintech companies are adopting robo-advisory due to advantages such as low operating costs, ease of scaling, and minimisation of human error and fraud. While the adoption is currently at a nascent stage, it is likely to become widespread in the near future. The potential benefits of financial inclusion that robo-advisory could usher in, make it particularly relevant and potent for the Indian market.

In the context of India, it is important to understand the various levels of complexity of the product landscape with respect to robo-advisory across various levels of complexity. Within the larger category of robo-advisory, there are platforms that simply offer a digital interface that provides investors with an

automated portfolio proposal with automatically selected funds or stocks. At the next level of complexity, there are robo-advisory services that are algorithm-driven, which offer automated execution and portfolio rebalancing services based on investment strategies that have been planned previously. At the highest level of complexity are robo-advisory services that are fully intelligent systems that self-learn and are driven by economic theories, without any significant human intervention. India currently has robo-advisories that are at the first two levels of complexity (Hon, 2019).

Although the number of individual investors has been increasing at a 10-year compound annual growth rate of 11 percent in India, access to wealth management services nevertheless remains limited. At the National Stock Exchange, a diverse set of participants are registered for varied product suite, but the total number of participants was 27.8 million in 2019 (Limaye, 2019) — which is a low number for a large market. Robo-advisory can make a large positive impact in this market, if it is guided in the right direction by stakeholders and allowed to serve consumers in a fair, equitable way. However, in order to do that, robo-advisories will have to first address certain key issues related to algorithmic decision-making.

Issues Surrounding Algorithmic Decision-Making and Trust Deficit

Algorithmic applications extend from providing shopping advice on e-commerce sites to performing medical diagnostics (Sharma, 2010). The wide and deep sectoral reach of ADS applications span a range of fields such as healthcare, retail, agriculture, manufacturing, transport, energy, smart cities or urban development, education and skilling, telecom, and of course, the software and IT industry itself. These industries use AI for applications that often directly affect consumers and users. Relying on complex mathematical and statistical models, such as deep neural networks, to recognise patterns and semantics from large volumes of data, these industries use AI for applications that often directly affect consumers and users.

While the ability to detect patterns has had transformative effects on these fields, these algorithms often deliver unexpected or counter-intuitive results. Lack of

data, biased data, privacy rules, use of wrong tools, irrelevant noisy variables and a number of other reasons could be the cause of this limitation. A machine learning system trained on man-made data is likely to pick up some unconscious biases already present in society (Garcia, 2016). Many such instances have come to light; language models learned from data have been shown to contain human-like biases (Angwin, et al., 2016). For instance, machine learning systems used for criminal risk assessment have been found to be biased against black people (Jeff Larson, 2016). Further, the use of propriety data of a user warrants secure protection by the said application.

To laypersons, AI solutions offer very little or no understanding of what happens in between the various stages of the process between the input of data and the output of results. Despite the unimaginable computing power and cost reductions, most developers emphasize more on incrementally improving the performance of AI systems according to a narrowly defined set of parameters and not on how the algorithms are achieving the requisite success.

Such challenges make it difficult to adopt and trust machine learning systems. In order to assess this risk and regulate algorithms, opening this machine learning 'black box' is necessary. By using algorithmic explainability techniques, designed to decipher algorithmic decision logic, pin point inconsistencies or cases of bias, the true potential of these algorithms could be unlocked. Due to the vast scope of applications and varying complexity, explainability is a challenging problem which has excited academicians and industrialists in the field of technology as well as public policy and resulted in astounding public interest in making these AI and ML algorithms explainable. This issue appears in popular press, industry practices, regulations, as well as hundreds of recent papers published in AI and related disciplines (Alejandro Barredo Arrieta, 2019).

Addressing Trust Deficit: Algorithm-centric and User-centric Approaches

Two broad approaches have emerged with respect to addressing trust deficit when it comes to artificial intelligence. The first approach attempts to instill human values in AI through a moral code. However, this approach has thrown up complex questions with respect to which value system can be used and how moral

and ethical frameworks would translate across cultural boundaries. Further, even if an AI-driven system was instilled with ethical values, the inability to feel emotional consequences in case of failure to abide by those values would continue to render them vulnerable to being bad moral actors. However, solutions like inverse reinforcement learning – where an AI is allowed to observe how people behave in various situations and understand what they value – are said to be showing promise. This approach also, brings with it a wider set of imponderables that are difficult to solve for in a quantifiable manner (IBM, n.d.).

The second approach is to increase transparency by making it easier for individuals to understand decisions being made by AI systems. In fact, industry leaders believe that the technology could get to a point within the next five years where an AI system can better explain why it is recommending certain outcomes to its users (IBM, n.d.). This is at the core of explainability in AI. The idea is to improve users' understanding of how the algorithm is producing results without opening of code or technical disclosures. The Local Interpretable Model-agnostic Explanations (LIME), for instance, is an algorithm that overcomes AI black boxes. However, with disclosure taking place over an extended period of time, another factor may come into play - behaviour change among users who may game the system by leveraging their understanding of the parameters at play.

Most research efforts looking into the explainability of AI takes an "algorithm centric" view, relying on "researchers' intuition of what constitutes a 'good' explanation" (Guidotti, 2018). Thus, the result is several varying definitions of an 'explanation'. Since the research is usually conducted by the machine learning and computer science communities, the focus is on explaining the inner workings of the algorithm. Despite emerging solutions to the black box problem, human intervention will be needed to interpret AI decisions.

Practically, in a system made for public use these explanations are not enough because they concern the lay users, who may not have a deep technical understanding of AI, but hold preconceptions of what constitutes useful explanations for decisions made in a familiar domain. This includes the data privacy policies that are generally based on fine print, making it unreliable and

inexplicable for the user. Additionally, since decisions and predictions made by AI are critical, trusting a neural network is difficult owing to complexity of work. Interpretation of data was easy because organisations that use highly sophisticated deep learning models could see acceptance and rejection of a user input but explainability was always a challenge. Therefore, in these practical applications, a **user centric process of determining a 'good' explanation** is required.

For example, one of the most popular approaches to explain a prediction made by a ML classifier, as dozens of XAI algorithms strive to do, is by listing the features with the highest weights contributing to a model's prediction (Marco Tulio Ribeiro, 2016). Another notable example is Google's What-If tool (James Wexler, 2019), an open-source application that allows practitioners to probe, visualise, and analyse machine learning systems, with minimal coding. A key aspect of the tool is its visualisation feature, allowing creation of intuitively-understandable customisable explanations.

Our contribution and approach to address these issues is through designing an explanation strategy that makes use of many existing explanation approaches and frameworks and testing it out with real users. We use the robo advisory use case to design user centric explanations of a specific high stakes' application with the goal of generating and judging high quality explanations in context. These explanations are user tested to understand specific as well as generalisable requirements of ADS users. Our objective is to analyse user thoughts on the usability of the explanations and the system. We also quantify user understanding and comprehension of different types of explanations. These requirements could provide insights leading to user centric policies that also aid the development process in the future.

Explainable Artificial Intelligence: Review of Literature and Discussions

Before delving into a discussion on the key issues surrounding XAI, it is useful to review the available research on the subject. This will help contextualise the discussions and help understand the rationale behind the approach we have adopted.

XAI has been around since the beginning of AI. There is a large body of research exploring its taxonomy, techniques, categories and evolution. The experiments conducted in this research borrow techniques from XAI, HCI-UX research and design. The purpose of this four-part section is to explore these techniques along with the justification for their relevance to this study. First, we define XAI and its goals and establish the growing importance of explanation from multidisciplinary perspectives. We also explore the popular trends in XAI and the evolution of techniques that provide algorithm-centric or user-centric views on explainability. Second, we define approaches to explanations of complex ADS, which we group into 3 broad categories, based on inherent explainability, scope and model accessibility. Third, we explain the background, logic, and goals of our user research. We also explain the types of information we wish to provide to the users through these explanations.

1. What is explainable AI?

The need for explainable AI was apparent as soon as AI systems were made usable. Research on XAI has been going on since the 1970's (Shane Mueller, 2019). From the figure below, it can be observed that the need for explanation has arisen in the last few years, as machine learning and deep net technologies have been expanding in scope, application, and reach (Abdul, 2018).

Explanation in AI

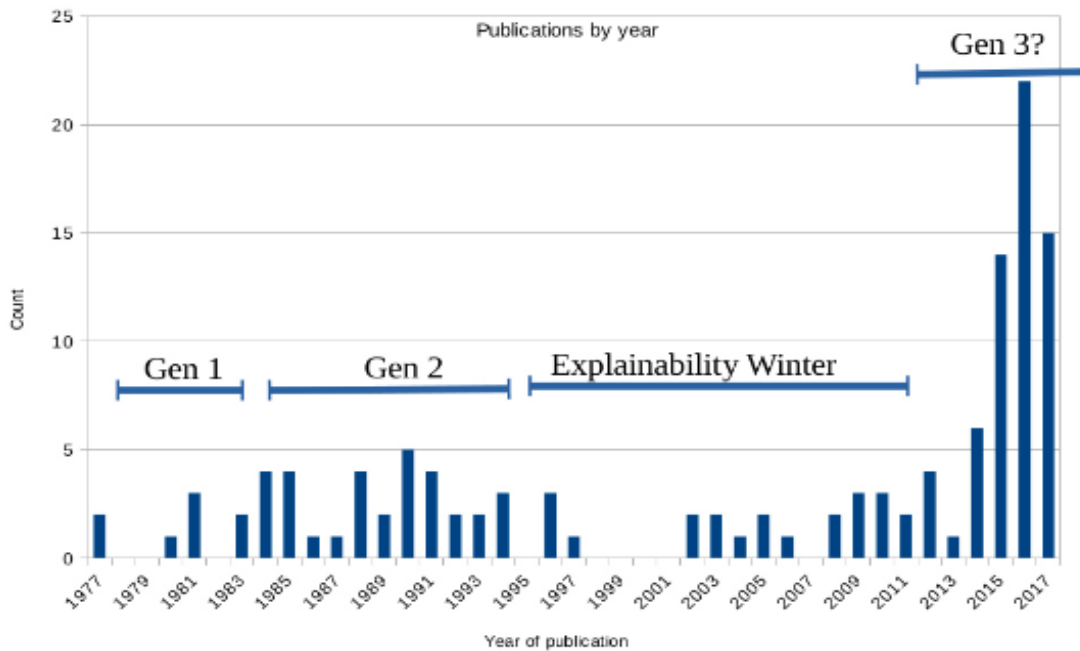


Figure 3. 1 Publications by year (1977-2017)

Figure 3.1 contains a histogram of the number of publications per year (1977 to 2017), identified in (Shane Mueller, 2019). Bibliometric analysis conducted by selecting publications with tags relevant to explanation in intelligent systems (AI, ML etc.). The DARPA project authors conducted a literature review of all publications tagged as intelligent system research and the timeline constructed above is based on this large collection of studies.

XAI is used to inspect an algorithmic system in order to understand the steps and models involved in making decisions, by asking and addressing questions surrounding why an AI system makes a specific prediction or decision. This need is further propelled by concerns surrounding the notorious 'black box' nature of AI algorithms with decision making unknown to the developers. The demand for XAI has increased calling for algorithmic regulation from the social sectors. This is critical to assuage certain ethical, policy and legal anxieties and concerns that surround ADS. Explainability is a way to gain assurance of the good quality of a system. XAI is beneficial to a diverse set of stake-holders involved in the ADS development and deployment process. XAI is especially beneficial to the users of AI systems, who require instructions on the purpose, appropriate usage and expected results in order to understand and trust these systems. For additional

literature on this topic, please refer to Appendix: What is XAI? What is the need for it?

2. Approaches to Algorithmic Explainability of Automated Decision Systems (ADS)

Approaches to explainability in XAI can be broadly divided into three categories: based on whether the algorithm used is inherently explainable in nature, the point of explanation generation in the development process (built in or post prediction explainability) and the type of view captured by the explanation.

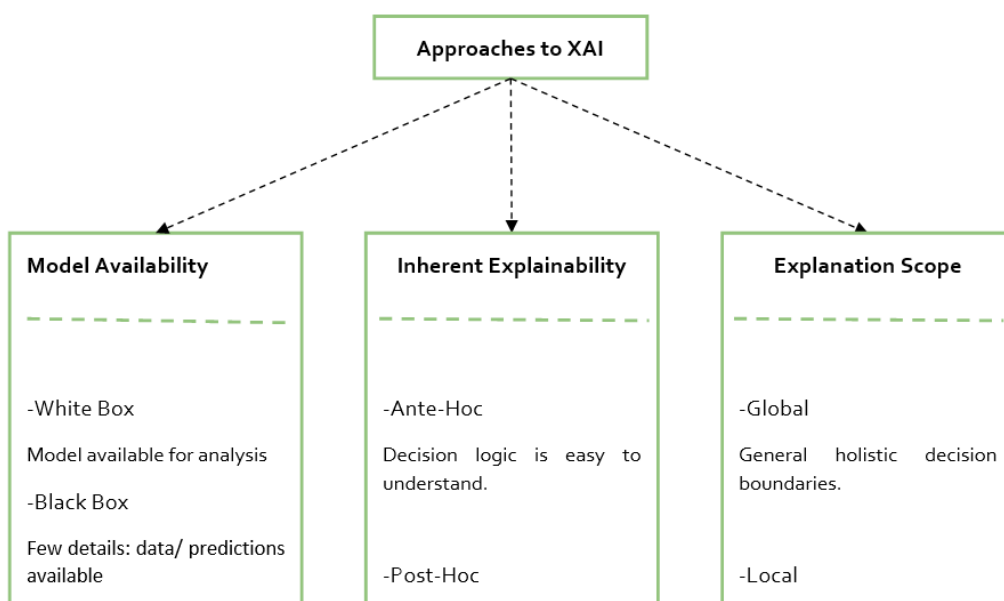


Figure 3. 2 Broad classification of XAI approaches

2.1 Explanations based on the inherent explainability of algorithms

First strategy of explainability depends on the complexity of algorithms. (Andreas Holzinger, 2017)

2.1.1 Ante-Hoc

Ante-hoc or before-this (event), are transparent by nature (glass box). Explainability is easier in parametric methods like linear models where feature contributions, effects, and relationships can be easily visualised and the contribution to a model's overall fit can be evaluated. This includes '**white box models**' such as Linear Regression, Decision Trees, K-Nearest Neighbours,

Generative Additive Models, Rule-based systems etc. These are easier to interpret, visualise and display since they progress similar to the human thought process. An example is the designed to be explainable RETAIN (Edward Choi, 2016) tool for application to Electronic Health Records (EHR) data. RETAIN achieves high accuracy while remaining clinically interpretable.

2.1.2 Post-Hoc

Post-Hoc or after- this (event/model) approaches provide an explanation for a specific solution of a '**black- box model**', For example, LIME (Marco Tulio Ribeiro, 2016), BETA, LRP (Avanti Shrikumar, 2017). Post-hoc explanations present a distinct approach to extracting information from already trained or learned models. While post-hoc interpretations often do not capture precisely how a model works, they confer useful information for practitioners and end-users of the system.

Generally, this method is used to explain complex neural network models, not inherently understandable. For example; non-parametric models, SVMs, multilayer neural networks, Bayesian inference systems and gaussian processes in ensemble models such as random forests. The inner workings of these models are difficult to understand and they do not provide an estimate of the importance of each feature on the model predictions, nor is it easy to understand how the different features interact.

To tackle these issues often a simplified human-understandable explanation is generated. These include visual explanations (through pie charts, bar graphs), textual reasoning and justifications, tabular explanations of patterns (structuring high dimensional data; directed graphs to explain causal loops and probabilistic inference), instance-based explanation (based on similarity measures also known as explanations by examples), simplification by surrogate models (simpler human-interpretable models are trained on the outcome data generated by complex models), and identifying relevant features based on the use case or domain.

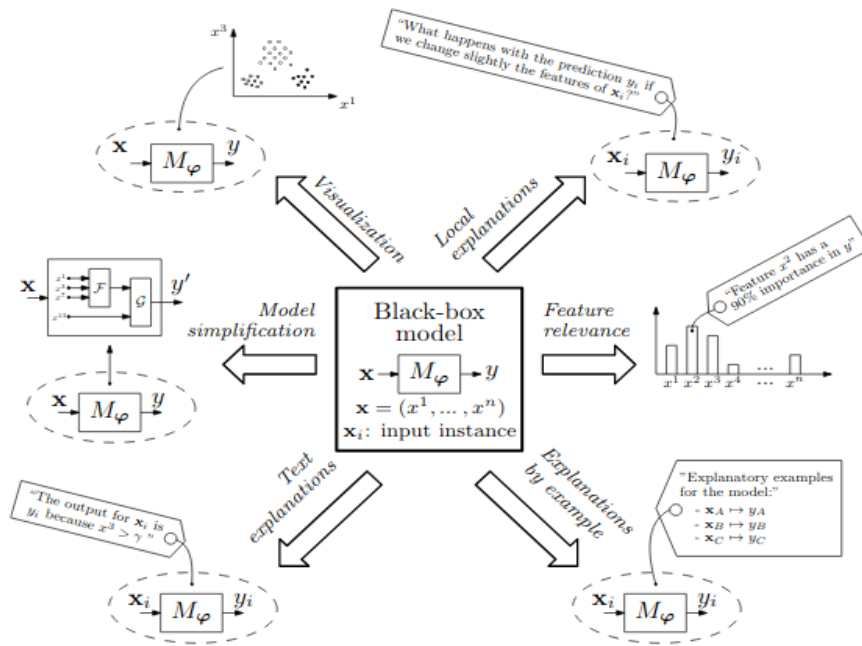


Figure 3.3 Post-hoc explainability approaches

Figure 3.3 contains the conceptual diagram designed by (Alejandro Barredo Arrieta, 2019), showing the different post-hoc explainability approaches.

2.2 Explanations based on the Availability of Models

This type of explanation depends on transparency of the model (Métayer, 2019).

2.2.1 White Box Explanations

This approach assumes that analysis of the ADS code is possible and the model is accessible. As this approach often requires re-running models with different samples of data. This approach enables explanations to ask and answers questions such as: What if I change my answer to another? Or, what result does my neighbour get?

A few examples of research on this approach are DeepLift by (Avanti Shrikumar, 2017) that uses 'Layer wise Relevance propagation' to check and compare the mathematical transformations input images go through in deep neural networks. FairML by (Kenneth Holstein, 2019) runs and reruns the available model by adding slightly perturbed input data, which helps understand the neighbourhood of data points and whether the model gives consistent results. Elvira (Carmen Lacave, 2000) is a system for the graphical explanation of Bayesian networks.

3.2.2 Black Box Explanations

This approach assumes that the analysis of ADS model is too complex or impossible. To get over this limitation, simpler surrogate models are used on complete or partial sets of input-output data. Explanations are constructed from interpreting the simpler model as well as observing the relationships between the inputs and outputs of the system. This approach can answer questions such as: Which features are important to the algorithm? How do my answers affect the outcome?

Examples of this category of approach include LIME (Marco Tulio Ribeiro, 2016) (Local Interpretable Model-agnostic Explanations), Anchor, TREPAN (Roberto Confalonieri, 2019), AdFischer and Sunlight. These approaches make use of surrogate/ simplified — representative models trained on the available data.

3.2.3 Constructive Explanations

The constructive approach is to design ADS taking explainability as a requirement. Two options are possible to achieve explainability by design:

Designing algorithms to explain their thought process, this could mean some accuracy – complexity trade-offs. (Been Kim, 2018) comes with a new technique to train neural networks, using concept activation networks. This would make the internal learning logic of the neural networks inherently understandable to humans such as RETAIN (Edward Choi, 2016), a tool designed to explain its decision with some restrictions applied on the algorithm.

The second approach, enhancing an accurate algorithm with explanation facilities so that it can generate, in addition to its nominal results (for example, classification), a faithful and intelligible explanation for these results. Neural Network designed with K-Nearest Neighbours in each layer; the neighbouring inputs near the relevant data point are generated as explanations, in order to check consistency (Tobias Plötz, 2018).

2.3 Explanation Strategies based on the Scope of Explanation

The third kind of explanations is based on the generalised-localised logic explained. Explanations could give an insight into the decision making of an entire

system or it could just explain the reasoning behind an individual prediction. (Claudia Biancotti, 2018), (Philippe Bracke, 2019)

2.3.1 Global Explanations

A global level explanation attempts to provide a holistic view of the system. These explanations can answer questions such as “How does the system make decisions?”, “What are the general patterns and relationships observed between input features and outcomes?”. A ‘global’ view of the system is definitely useful but it does not reveal the intricacies of recommendations given by the ADS. Global views are relatively more human understandable, depending on the complexity of the model. For instance, for a support vector machine, classification is carried out by drawing multiple hyperplanes across a multi-dimensional space. Beyond a point, it is difficult even for developers to understand the intricate divisions. Thus, global explanations provide a very useful view of the system. For example: Showing feature attribution, Partial Dependence Plots, Individual Contribution Plots.

2.3.2 Local Explanations/ Justifications

Local explanations try to answer questions regarding specific predictions; why was a particular prediction made? What were the influences of different features while making that particular prediction? As mentioned above, in our case, local explanations will help explain why a particular user is given specific advice by the model or ADS. For example, LIME (Marco Tulio Ribeiro, 2016), that stands for ‘local interpretable model agnostic explanations’, explains the logic behind classification for each prediction.

From given examples, we can understand that these approaches are **not mutually exclusive** and contain many overlaps. Each has its own scope-based merits and are broadly classified based on model availability and accessibility. In this study, we cover white and black box explanations, generated ante or post hoc to explain both local and global views. We wish to study the effectiveness of these explanations on user comprehension and usability of the system.

3. Measuring Effectiveness of Explanations through User Studies

The previous section described techniques to generate explanations. The next pressing question is: what makes these explanations 'good'? And, how to determine the effectiveness of generated explanations? In the context of our research, the 'correctness' of user-centric explanations implies user understandability and 'goodness' is determined by explanation usability. Easy to understand explanations could also affect users' trust in the accuracy and accountability of the system and its recommendations. For additional literature on this topic, please refer to Appendix: Literature review on measuring effectiveness of explanations.

Important considerations while building robo advisory explanations were taken from the review. First, explanations are usually short answers to 'why' questions. Second, good explanations are contrastive or relative. Third, explanations rarely consist of an actual and complete cause of an event. Explanations are meant to be a social transfer knowledge, presented as part of a conversation or interaction, and are thus presented relative to the explainer's beliefs about the explainee's beliefs. Fourth, causal relationships convey information in a human understandable manner as opposed to referring to probabilities or statistical relationships in explanation.

Therefore, the explanations provided in our system should cover all key concepts covered above: causal reasoning (identifying cause and effect relationships, or, does A cause a change in B?), abductive inference (start with results and then understand the logic behind decisions), counterfactual reasoning (calculating the changes in results if inputs are changed, or, If A then B? type questions), contrastive reasoning (comparison with alternative explanations, or, Why A not B?) and understanding of complex systems (interactions and narratives to create mental models to trace user thinking).

Users are given a replica robo-advisory system to experiment with. Explanations are intended to satisfy user requirements listed below:

1. **Information requirements:** required knowledge to provide an adequate explanation.
2. **Information access:** justifications, what information the explainer has to give the explanation such as the causes, the desires, etc.
3. **Pragmatic goals:** refers to the goal of the explanation, such as transferring knowledge to the explainee, making an actor look irrational, or generating trust with the explainee.
4. **Functional capacities:** each explanatory tool has functional capacities that constrain or dictate what goals can be achieved with that tool.

Once the users are satisfied with an understanding of the system procedure and goals, they turn towards the explanation of the algorithms used in the process. Whether users comprehend these explanations is examined through quantitative survey questions. Appendix: Table 2 (User Survey Contents) shows all the questions used in the survey.

Research Questions and Goals of our user study

Our research aims to contribute both theoretically and empirically to enhance human understanding of complex algorithmic systems. The previous sections delve into the definition of Explainable AI, its need, growth, evolutionary trends, and the points of segmentation considered while generating explanations; namely model complexity, availability and scope. We also explain the emergence of robo-advisory applications and the trust deficit due to a lack of transparency that hinders their potential to democratise finance. With our experiment, we wish to directly assess user interpretations of the explanations of a robo-advisory application and their effect on the usability of the application. Ultimately, this contributes to a standardised framework of explanations to broaden user understanding of complex decision-making algorithms. A summary of the research questions addressed through our study is given below.

1. How are users' thoughts and opinions influenced by explanations, especially quantifying changes in user trust in the advice and recommendations of the robo-advisory system?
2. How do explanations affect the usability of the robo-advisory complex algorithmic system as a whole?
3. How do explanations and consequently user opinions vary based on the complexity of algorithm and nature of explanations? (White box vs. black box explanations)
4. How does user perception of explanations system usability change in context of demographic group membership? For example, users from different age groups, risk categories, backgrounds, prior robo-advisory or investment knowledge, etc.
5. How could the information gathered on user comprehension, opinions, and usability of explanations contribute towards the broader picture of generic guidelines for innovative inclusion? In order to aid developers and designers of ADS to help regulators better frame policies in the future.

In the experiment, designed to address our research questions, we first design a system made of complex algorithms equipped with different types of

explanations. After this, we test the explainability of the system with a survey posed to active users. Designed to evaluate the usability and interpretability of these explanations. We gather user thoughts and opinions in order to analyse user behaviour (such as frustration, trust, etc.). Drawing from the broader implications of the survey results, we analyse the effectiveness of our strategy. Our ultimate aim is to contribute towards an effective and generalisable framework of explanations that are human understandable.

Methodology & Experiment Design

The primary aim of the study is to analyse the effectiveness of user-centric explanations in conveying the decision-making logic of complex algorithmic systems. The previous sections reviewed the literature that set the stage for our experiment. Detailing the multi-perspective approaches towards system explanations, the interdisciplinary motivations behind explanations, and evolutionary trends followed by similar research. This section is broadly divided into two categories. First, materials and tasks for the experiment which includes designing the RA systems and their explanations. Second, the design of the user study experiment, content and structure of user surveys, sampling and recruiting users.

1 Scope of our experiment - human centric XAI strategies.

This experiment follows a **hybrid approach of application and human grounded evaluation** of explanations (Finale Doshi-Velez, 2017). Broadly, the tactics followed are:

- Benchmark/baseline determination: The target demographics of the system (novice users) and users already familiar with the system (advanced users) or domain experts are selected as experiment participants.
- These users experiment with a simplified version of the original system containing all the basic functionalities, equipped with user-centric and human understandable explanations.
- These users will be asked quantitative and qualitative questions to evaluate the usability, understandability of various explanation strategies, and the effect of explanations on usability and trust in the system as a whole.

This approach to evaluate the quality of explanations is selected from one of 3 commonly studied strategies presented in Figure 4. While the alternatives use proxy humans or tasks, application-grounded evaluation follows an interdisciplinary approach of using research techniques from human-computer

interaction, data analysis, and visualisation along with XAI principles. The real task used for the experiment is a Robo advisory application that recommends mutual funds for investment.

	Application-grounded Evaluation	Real Humans	Real Tasks
	Approach used in this study. Tasks and experiments are designed with a real application in mind. Other examples: (Pedro Antunes, 2012), (Jonathan Lazar, 2010)	User centric explanations of system logic tested on real humans.	Replica robo advisory system designed for users to experiment.
	Human-grounded Evaluation	Real Humans	Simple Tasks
	Conducting simpler human-subject experiments that maintain the essence of the target application. Examples: (Been Kim, 2013), (Himabindu Lakkaraju, 2016)	Asks for user opinions directly rather than gathering them through tasks.	Tests more general/ abstract/ qualitative notions.
	Functionally-grounded Evaluation	No Real Humans	Proxy Tasks
	Uses some formal definition of interpretability as a proxy for explanation quality.	No human experiments conducted.	Similar instances are evaluated rather than the direct task.

Figure 3. 4 Experiment design

Figure 3.4 explains the approach taken in this experiment and highlights taxonomy with alternatively used approaches to interpretability studies, based on research presented in (Finale Doshi-Velez, 2017)

2 Describing our Experiment

The current section intends to expound the experiment itself. As explained in the section above, the experiment provides users with a complex application task, then presents explanations of this system. A set of qualitative and quantitative questions measure user interpretation of these explanations and ultimately their effectiveness. These questions are designed to capture user comprehension and explanation usability by satisfying user needs (detailed in the literature review). These users need for information on the system covers: abductive inference, causal, counterfactual and comparative explanations. The application task or use case is a mutual fund recommending Robo advisory application. The system profiles each user into risk categories and recommends mutual funds customised to user preferences gathered from a series of questions. Complex machine

learning algorithms calculate user and fund risk and then select and allocate ideal funds for investment.

First, we describe the complex algorithmic system constructed as the "task" of the experiment: a small Robo advisory system conducting risk determination and asset allocation tasks. We dive deep into the construction, implementation, and decision-making logic of this intelligent application. Second, we describe our framework for explainability covering multi-perspective explanations of the Robo advisory complex system. Our framework covers two types of cases divided on complexity and transparency: white box case (the ideal scenario of complete algorithmic model availability, where explanations are generated without any limitations) and a black box case (a more pragmatic scenario, considering the limitations imposed on the availability of model due to lack of transparency in the model design due to increased complexity or confidentiality/ protection of intellectual property etc).

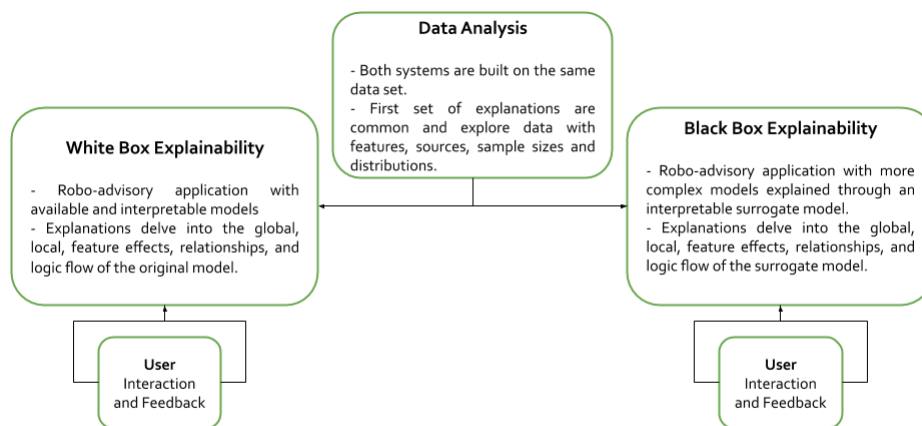


Figure 3. 5 Abstract procedure for the task in the experiment

3 Replica Robo-Advisory System

The task given to participants of the user study represents a real-life, fintech application of robo-advisory applications. This application aims to recommend mutual funds to potential and current investors. This use case is an ideal example demonstrating the need for explanations in complex algorithmic systems. This section contains details on the design and development of this robo-advisory

application. Two decisions made by a robo-advisory application directly concerning users are risk profiling and asset allocation. Understanding the rationale behind these decisions is crucial. Therefore, we focus on generating explanations to convey this decision-making logic. Explainability is appreciated and essential in the financial sector since the impact of an incorrect risk profiling and recommendation could result in the loss of users and assets. An additional benefit for the fintech sector is a sophisticated set of users who are accustomed to analysing complex information usually visual and digital in nature. Consequently, this makes user selection easier. This study also follows our previous research on generating explanations for Robo advisors (Krishnan, et al., 2020) that explores alternative explanation strategies for black box algorithms.

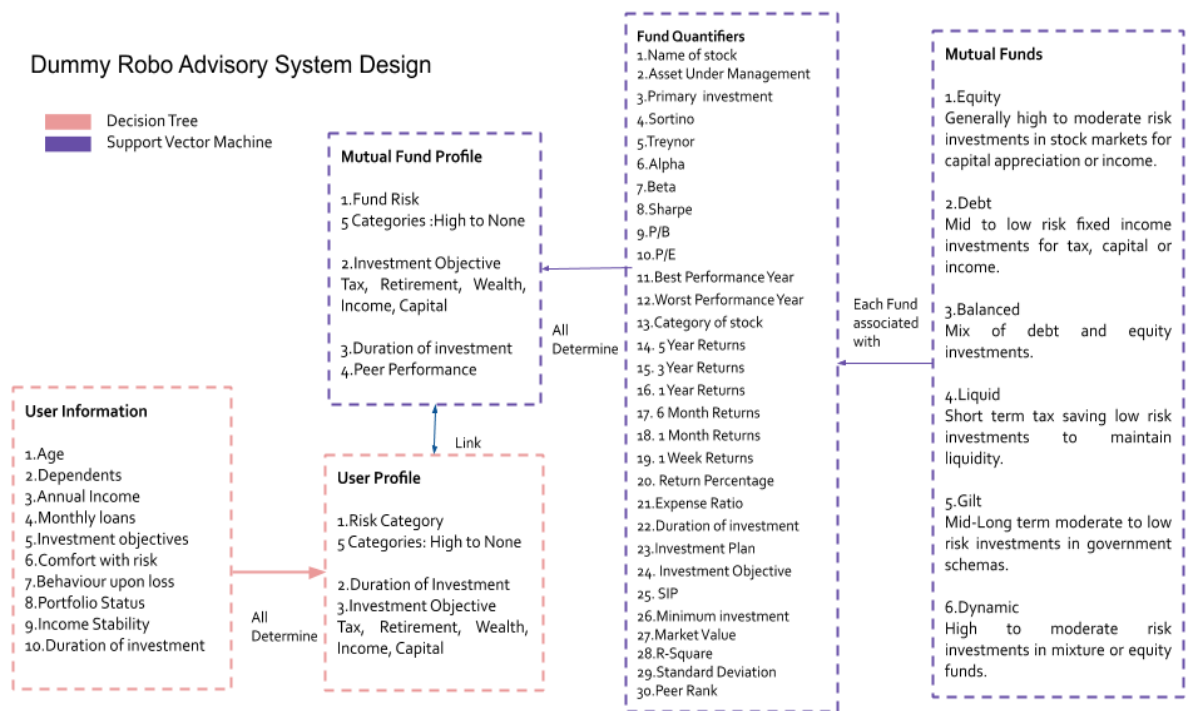


Figure 3.6 Robo advisory system designed for the user study

As Figure 3.6 shows, user profiles and mutual fund profiles are determined using supervised classification algorithms. Customised recommendations are made based on commonalities. Three algorithms work together to provide mutual fund recommendations to users; the first determines user risk profiles (pink), while the second constructs a mutual fund risk profile (purple), and the third recommends

mutual funds based on matching these two profiles through a set of different parameters (risk, objective, duration and amount of investment).

Two different machine learning models form the heart of the system (refer to Figure 5). The first model creates a user profile after gathering user preferences, capturing useful information regarding investor objectives, limitations. But its primary purpose is to assign a risk category to each user. The second creates a mutual fund profile by classifying mutual funds into the same divisions of risk, based on a large amount of historical data gathered on 30 variable factors associated with each fund. This model also captures specific details regarding investment required, fund objectives, and duration for an adequate return. The models have been selected based on an evaluation of their performance metrics (Krishnan, et al., 2020). Both these models differ in complexity. In the white box system the decision tree is inherently explainable while the support vector machine is more complex. In the black box system, both algorithms used are support vector machines and their explanations are generated through inherently explainable algorithms acting as surrogate models. The difference in the complexity of the white box algorithms provides an additional point of view to evaluate user understanding. As explained in the Literature Review section, an increase in the complexity of algorithms is reflected in its explanations. Due to the increase of features and complexity between the models, in this case, the cause-and-effect reasoning is difficult to convey through the explanations. The survey results show the consequent difference in user comprehension. For details on algorithms, please refer to Appendix: Algorithm Details.

To sum up, the robo-advisory system gathers user preferences through a questionnaire. Based on these preferences a risk profile is created for each user. This profile is matched with many mutual fund risk profiles using a content-based recommendation system. These fund profiles are created using another algorithm. Therefore, our explanation framework should cover aspects of each one of these algorithms.

3.1 User Profile and Preferences

A set of questions meant to gather personal details from user preferences begin the Robo advisory process. This creates a unique user profile, used to recommend customized stock portfolios. (Krishnan, et al., 2020) provide a list of questions commonly used across Robo advisory applications such as (PayTM money, 2010), (ETMoney, n.d.), (Tavaga, 2018), etc. Each one of the questions presented to users are shown in the appendix (refer to Appendix: User Risk Questionnaire). In our system, each answer choice has a varying magnitude of impact on the final risk category assigned to each user (from low risk to high risk). The model evaluates each choice selected by a user to create a risk profile. For example, a user that selects age as 15 to 35 and objective as monthly income will be assigned a different risk category than a user with the same age and different objective or income, etc. A final user profile is then assigned to each user, focusing on risk scores, the objective and duration of investment, along with any financial limitations. Since these questions are of a personal nature, user anonymity is maintained by encrypting the names and details of each user on the online database.

3.2 Mutual Fund Performance, Risk, and Recommendations

After user risk calculation, another algorithm creates a risk profile for each mutual fund. While fund matching and user risk calculation is done real-time, the fund profiles are precalculated and stored in the system. The fund and user profiles are matched using a content-based recommendation system that considers user risk, investment objectives, financial capabilities, and preferred duration of the investment. Matching the user and fund risk are the most pivotal deciders for recommendations.

This risk associated with each fund is determined through thirty variables associated with each fund encompass historical performance, returns, expense ratios, and performance indexes (refer to Figure 6). These quantifiers include fund specific parameters to provide insight into the performance and behaviour of a particular fund; Alpha, Beta, Net Asset Value, Treynor, Sortino, returns, etc, and peer performance parameters to compare the fund with similar others; P/E and

P/B ratios, asset under management, best and worst performance. Additionally, fund comparison with benchmark or index funds are considered. These quantifiers are based on parameters followed by seasoned investors to understand multiple aspects of fund performance, such as: Investment objectives of the fund (does the objective match mine?), Monthly capital needed to invest in the fund. Does the fund fulfil investment objectives? Growth plans of the fund. Duration till return on investment. The necessary and unnecessary risks taken by fund managers (whether this risk matches my risk comfort). Comparative peer performance (are these the best funds in this category?). These choices are made based on a study of many expert web sources, which classify funds into risk categories using the same quantifiers as indicators (AMFI), (Value Research), (Money Control), (MutualFundIndia). Broadly, mutual funds can be Equity, Balanced, Debt, Liquid, Gilt, Dynamic, ETFs, Fund of Funds, and Specialty. Each one of these options appeal to a different investment objective, and the robo-advisor can recommend a fund from each division. Refer to Appendix: Mutual Fund Quantifiers and Types for details regarding fund and variable selection.

Our system contains a supervised learning algorithm trained on a small dataset of mutual funds with all the above quantifiers as dimensions. The labels for supervised learning are gathered from multiple popular and reliable sources such as (AMFI), (MutualFundIndia). This pretrained model is then used on a large list of mutual funds to assign a risk category to each. After this process, each fund is stored with a profile highlighting fund risk, duration of investment, returns, objectives, growth plans, and initial investments. This algorithm performs the task with sufficient accuracy, preciseness, and recall.

4. Generating Explanations

To convey the rationale behind these algorithmic decisions, we generate explanations for algorithms used in the robo-advisory system. In our study the two explanation systems are separated by a major point of segmentation: the extent to which algorithms are accessible and available. This could either depend on the opacity of the algorithm or to fulfil a need for abstraction due to confidentiality of intellectual property. We introduce our explanation framework

through two abstractions of explanations systems, a white and black box. White box explanations present a global and local, ante, and post-hoc analysis, constructed in the ideal scenario of complete model availability. While black box explanations assume little to no model availability and provide a global post-hoc view of algorithm analysis. For a view of the entire system refer to Appendix: Complete experiment overview of the white box system

4.1. Exploratory Data Analysis

Across both these systems a common explanation is of the essential details of data used by algorithms with the objective of providing users with insight into structures and sources of data. The key factors that capture the essence of data is its type (nominal, categorical, etc.), quantity, source (well-known/reliable, original/processed, etc.), and content (dimensions, features, consistency, bias, etc.). These essential indicators of quality regarding the model and data are derived from studies with hopes to standardise the process of explaining tabular data, such as (Timnit Gebru, 2018), (Sarah Holland, 2018). In the robo advisor, potential investors are primarily concerned with the dimensions or variables used, and sources or origin of data. We also include information on distribution and stratified nature of sampling used in both algorithms. Therefore, explanations include sample distribution, variable information, sources, and types. These are conveyed to users through other white and black box explanations as well.

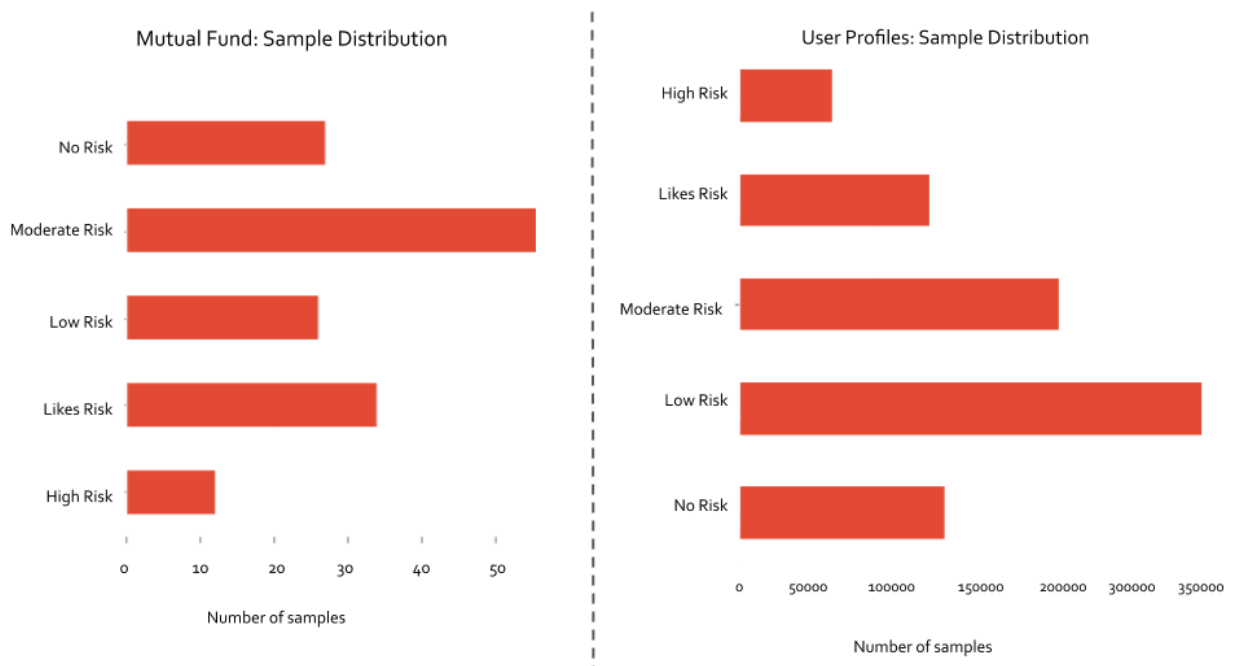


Figure 3.7 Data distribution of samples

Figure 3.7 is presented to users as an explanation of the data distribution of samples used for both algorithms.

4.2. White Box Explanations

White box explanations are meant to precisely and accurately convey logic used by the system, even user-specific questions such as “Why did I get this particular recommendation? Why not another one?”, “What did my neighbour get?”, “Who else received the same recommendations?”, “What happens if I change my answers?” etc. These explanations cover everything from the importance given to variables and their impact or transformations, decision boundaries of the algorithms, establishing benchmarks to compare predictions, and in some cases transparent flow of decision logic. Our selection of the white box explanations is based on our objective to enhance user understanding by satisfying counterfactual, abductive, contrastive, causal reasoning.

Therefore, our explanations include model details and descriptions mixed with data details, variable importance on general decisions (global) that affect every user as well as user-specific triggers (localised), interactions and behaviours of these variables with each other, and the final risk output. Additionally, white box

explanations explore group comparisons between two risk classes as well as transparent decision boundaries.

Although white box explanations combined with the data exploration seem like the answer to all explainability issues, this is often not the case. As the complexity of the algorithm increases, the complexity of variables and their transformations also increases. Not just for users but also domain experts and developers. Therefore, to some extent, explanation strategies need to depend on the nature and complexity of the algorithm. While the information needs of users are constant the design and choice of conveying the explanation changes based on the domain and model. We test our explanation strategy using a more complex algorithm through both white and black box cases of model explainability.

4.2.1 Model Details and Descriptions

Traditionally, functional machine learning systems are required to possess train, test, and validate performance details (AUM, Recall, Precision, F1-scores are general metrics present in such documents). Model details disclosed in this study have been derived from many frameworks that address this need (Timnit Gebru, 2018), (Margaret Mitchell, 2019), etc.

Apart from the metrics and data details, the explanations disclose the context in which models are intended for use, performance evaluation procedures, sampling strategies, and abstract decision logic followed by the model.

Model Function	Assigning a risk category to mutual funds to match and recommend users. This algorithm uses data gathered for a few mutual funds, to classify all mutual funds into optimal risk categories according to variables associated with each fund.			
Model Details	Support Vector Machine with polynomial kernel.			
Model Decision Logic	Model is presented with a set of data and categories of each fund. It learns from this data and draws complex convolutional boundaries around multiple values of variables belonging to one category. Numerous boundaries are defined in this multi-dimensional space and this logic is used to classify new funds.			
Model Output	One of 5 risk categories: No risk, Low risk, Moderate risk, Likes risk, High risk			
Sensitive Variables	No sensitive variables used.			
Performance Metrics	Average Recall : 88%, Average Precision : 80%, Accuracy : 90%			
Sources	Dataset and variables gathered from AMFI India, Money Control, Value Research, Mutual fund india.			
Sampling Strategy	Stratified sampling: proportionally equal sampling of data			
Variables/ Columns used to determine fund goals and risk	Name of stock Asset Under Management Primary investment (equity/debt).. Sortino Treyner Alpha Beta	P/B P/E Sharpe Best Performance Year Worst Performance Year Category of stock 5 Year Returns 3 Year Returns	Expense Ratio Term/ Duration of investment Investment Plan Investment Objective SIP Minimum investment Market Value	R-Square Standard Deviation Peer Rank 1 Month Returns 1 Week Returns Return Percentage 1 Year Returns 6 Month Returns

Figure 3. 8 Model details for mutual fund risk profiling algorithm

Figure 3.8 shows the model details for mutual fund risk profiling algorithm. It includes the intended usage and function of the model, decision logic, expected outcomes. The data details cover the variables used, sources, and sampling strategies. Commonly used performance metrics of machine learning models are also provided for user perusal (precision of the learning algorithm, recall, etc.).

Model Function	Assigning a risk category to users based on their preferences. This algorithm uses data gathered from past users, to classify new users into optimal risk categories according to their selected preferences.
Model Type	Decision Trees
Model Decision Logic	Tree like hierarchical decision boundaries are defined by repetitive linear divisions of data around every variable value. Boundaries made based on impact features make on model decisions. This learnt logic is then followed for future user classification.
Model Output	One of 5 risk categories: No risk, Low risk, Moderate risk, Likes risk, High risk
Sensitive Variables	Age could be considered a sensitive variables.
Performance Metrics	Average Recall : 96%, Average Precision : 96%, Accuracy : 97%
Sources	Dataset and variables gathered from PayTM Money, Tavaga, EZMoney, Betterment, Wealthfront.
Sampling Strategy	Stratified sampling: proportionally equal sampling of data
Variables/ Columns used to determine fund goals and risk	Age, Dependents, Income Stability, Annual Income, Comfort with taking risk, Monthly loans, Investment Objectives, Duration of investment, Behaviour after loss, Portfolio status

Figure 3. 9 Model details for user risk profiling algorithm

Figure 3.9 contains the model details for user risk profiling algorithm. Metrics and model types are clearly disclosed. An extensive list of sources is provided.

4.2.2 Feature Importance

This type of explanation is chosen to aid users to form mental models by conveying information about what goes 'inside' a model. Variables or features of the dataset are the inputs received by the algorithm. Decision logic is based on the impact and transformations caused by mathematical variations of these features (the effect is seen either directly or indirectly). Therefore, knowing how important each feature is to the model provides users with a highly compressed, global insight into model behaviour. Features that are important to a single individual prediction impart a personalized or localised insight into the effect of their choices for specific predictions.

4.2.3 Global Feature Importance

These explanations highlight importance of each feature with respect to each class. In robo advisors, users are classified into one of five risk classes determined through features selected by users such as age, annual income, dependents, etc. Each selected feature contributes towards the decision-making logic, positively or negatively, pushing the user into a certain risk category. Feature importance is calculated using Shapely values, by permuting each feature and measuring the average effect on the outcome (Scott M. Lundberg, 2017). A similar approach is adopted for mutual funds algorithm explanations.

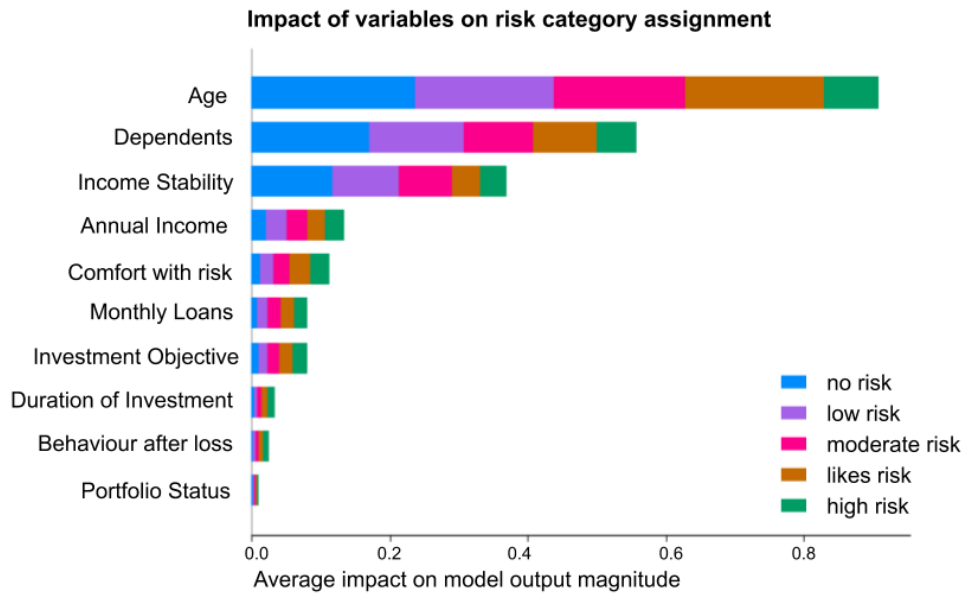


Figure 3.10 Impact of variables on risk category assignment

Figure 3.10 shows the global user risk feature importance, calculated by average marginal contribution of a feature value over all possible permutations. This explanation conveys the importance of each feature towards a prediction, the supervised algorithm assigns importance to each feature based on a labelled set presented to it. Details on the set is covered in the Replica Robo Advisory section. All the features used to determine user risk are listed in descending order of impact. According to the explanation generated for a sample user, the age of the user is the most important deciding factor, contribution is slightly higher to 'no risk' (~0.22) and 'likes risk' (0.60-0.82~0.22), but overall, almost equally to every class. Additionally, since all features contribute the least to the 'high risk' class, the algorithm classifies an instance in this class if it fails to fall in the boundaries of other risk categories.

Due to fewer variables used to determine user risk, each plays an essential role in classification. The model has observed nuances of each variable choice and is making use of every available variable while making a class decision. Since there is a significant increase in the number of variables used to determine mutual fund risk, this trend will change in the next figure.

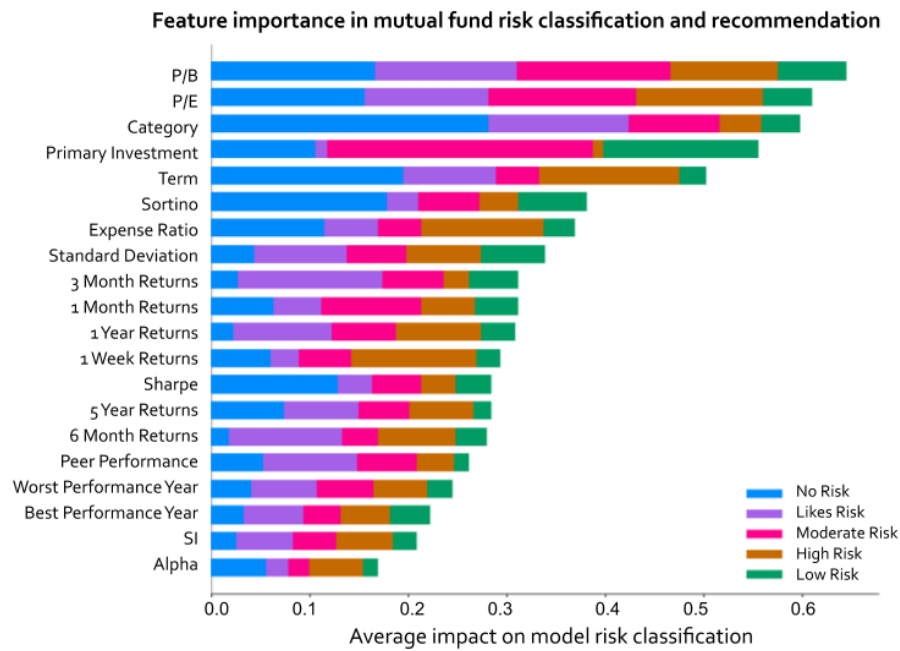


Figure 3.11 Feature importance in mutual fund risk classification and recommendation

In Figure 3.11, we can observe a greater extent of variation in global mutual fund risk feature importance. P/B and P/E ratios seem to draw the primary classification boundaries, causing an approximately equal impact in all classes. The model finds the 'category' of the funds an important indicator of 'no risk', that is, whether a fund is not risky can be determined based on its category alone, to a high degree of confidence. Similarly, we can observe each variable can impact risk decisions differently. For instance, 'primary investment' matters to the algorithm while deciding between 'no risk', 'moderate risk' and 'high risk'. But it cannot help while deciding whether the fund 'likes risk' or is 'high risk', for this the model will probably turn to 'expense ratio' or 'term' to measure 'high risk', or it will look at the '3-month return', 'P/B' or 'category' again for 'likes risk'.

To sum up, global feature importance provides insight into the 'thought' process an algorithm goes through when it encounters new values. Further, this represents 'learned' rules or behaviour, acquired during the training phase. Although the explanation helps users create mental models and context for understanding, it lacks depth. One cannot decipher whether the impact made on the class is positive or negative and as each variable can take different values (variable 'age' can be 15-35 or 35-55 or 55+), each internal segment can cause varying magnitudes of positive/negative impact on the outcome.

4.2.4 Local Feature Importance

'Local' implies the localised impact of the variable values for a single individual, the current user. Each user is shown how their choices have contributed directly to the risk decision. The plot below shows the positive and negative impacts of your answers on the algorithm's decisions to put you in **Low Risk**. Since the positive impacts outweigh the negatives, the algorithm will assign you this category specifically. The varying magnitudes of the impacts show that some features are more important to the algorithm than the others.

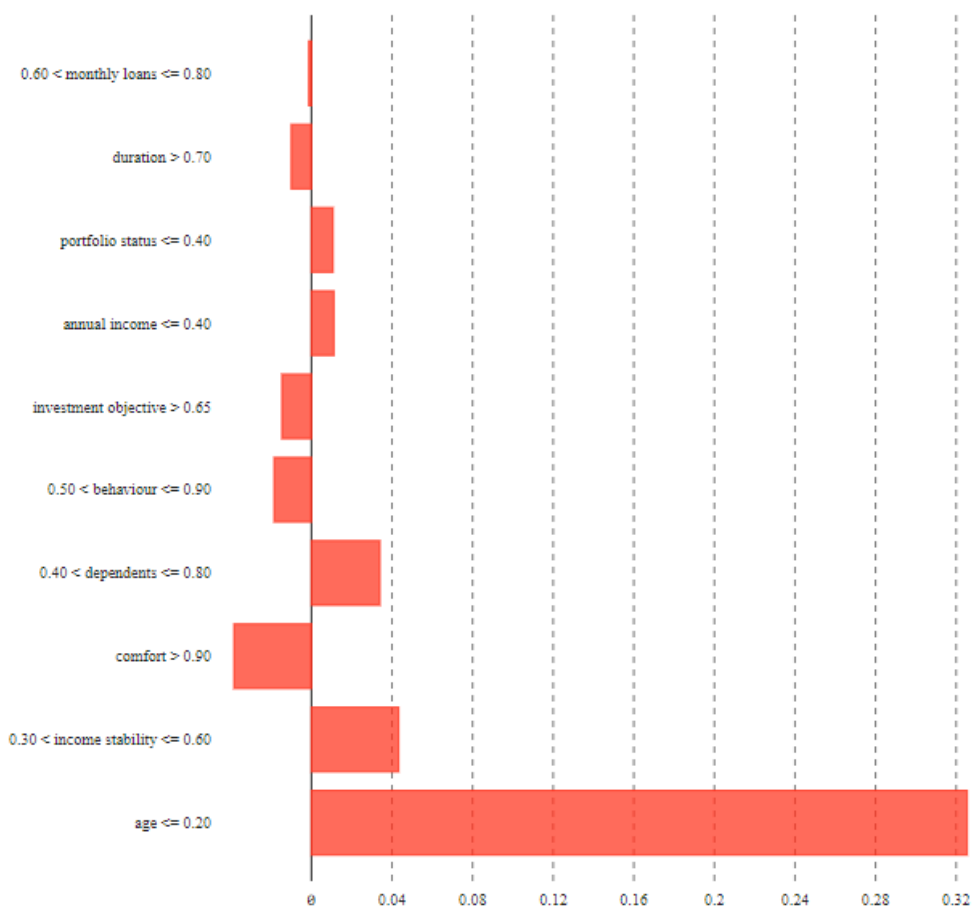


Figure 3.12 Risk category explanations given to users

Figure 3.12 illustrates the explanation of the risk category given to the users is specified above, along with the positive and negative impacts of the answers towards the class decisions. Locally interpretable explanations (Marco Tulio Ribeiro, 2016) are dynamically generated based on the answers selected by the user in the Robo advisory questionnaire. The user above, a married 55+ year old, wishes to invest his moderately stable, low income for 3-5 years. These factors

contribute positively to the 'low risk' classification. But he is comfortable with risk and wants a monthly income from his investments and both these factors try to push him into the 'moderate risk' class, therefore contributing negatively towards the 'low risk' class. Overall, the positives outweigh the negatives as the user is classified as 'low risk'.

4.2.5 Feature Effects and Behaviours

These explanations aim to overcome the 'depth' limitations faced by the global feature importance by conveying how values of each feature affect the outcome class and whether this effect is positive or negative. Each class has a unique combination of variable values and features used of distinction. Moreover, these explanations allow a global comparative view of feature behaviour learnt by the model (young age and low stability vs older age and high stability etc.). The *interaction* between features is defined by the change in the prediction that occurs by varying the features after considering the individual feature effects. These explanations help provide insight into the trends captured and convey *causal and counterfactual logic* used by the model.

The user risk calculation makes use of many categorical variables, depicted in Figure 3.13 is the colour assigned to these variable choices in table 3.1.

Questions	Options and values
Age	18-35 36-55 55+
Dependents	No one Spouse only Spouse and children Parents only Parents, spouse and children
Income Stability	High Moderate-High Moderate Moderate-Low Low
Annual Income	Below 1 Lakh 1-5 Lakh 5-10 Lakh 10-25 Lakh Above 25 Lakhs
Comfort with taking risk	Strongly Agree Agree Neutral Disagree Strongly Disagree
Monthly loans	None Upto 20% 20-30% 30-40% Above 50%
Investment Objectives	Wealth Creation Monthly Income Capital Preservation Retirement Planning Tax Saving
Duration of investment	3-5 Years Less than 1 1-3 Years 5-10 Years 10+ Years
Behaviour after loss	Invest More Do Nothing Sell after market recovers Sell and use FDs Sell and Preserve Cash
Portfolio status	Stocks Mutual Funds Bonds/Debt Savings/FDs Real Estate

Figure 3. 13 Categorical variables used in user risk calculation

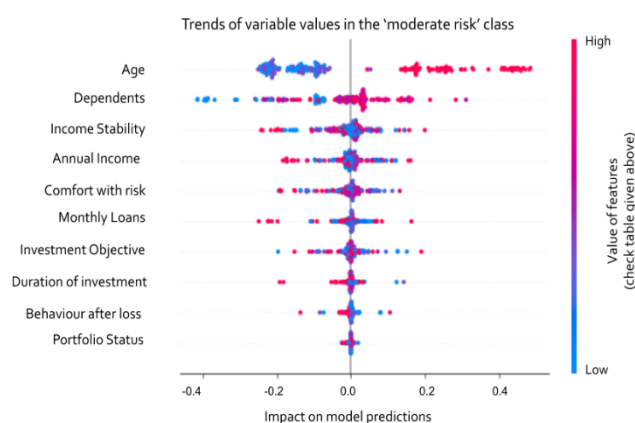
Figure 3.13 provides context for the feature effect diagrams. The questions asked by Robo-Advisor and the value range associated with the answers is shown above. The colours assigned to these variables are also used in the following risk trend diagrams. The colours are assigned based on the decreasing (red to blue) weights assigned to each variable value by the model. Machine learning systems recognise patterns in data by assigning each variable a numerical value. As we move from red to blue the risk-taking tendency decreases. Therefore, a person with every variable value blue will be highly risk averse while a person with red variable values will be a risk taker. The colours are designed to increase human understandability and the readability of information compared to its original numerical format.

The risk trend diagrams show the behaviour and effect of a sample of variable values in a specific class, magnitude and positive or negative. This is done using a few instances randomly of samples from the data. The jitter of the instances shows concentration of values. The interaction of variables can be measured by following each general trend (“What happens for a certain age value and a certain dependence value?” etc.).

Table 3. 1 General trends across classes for each feature

Trends in User Risk Classification¹⁵

Description

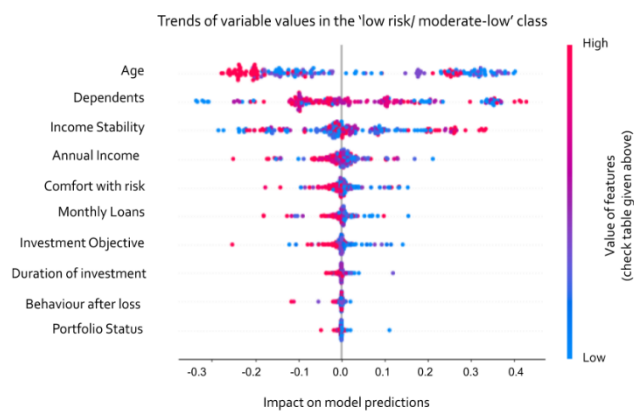


Moderate Risk

There is a clear distinction of the impact of different age values. There are more young (red: 18-35) users than others in the moderate risk class. If a user is 35+ the probability of belonging to this class is reduced. Looking at the

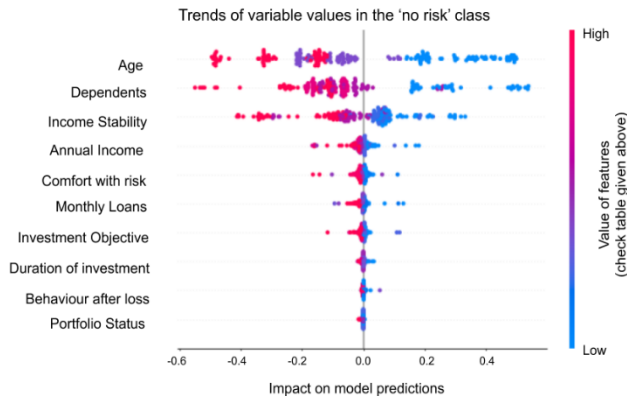
¹⁵ The colours represent the category score in a feature. The distribution of points in each feature corresponds to the distribution of data points in the sample, and also indicates the extent of negative or positive influence.

trend of dependents, users in this risk category have 0 to 2 dependents. All other variable values seem to have no impact on the classification. Therefore, dependents and age matter most in this class. Additionally, one could imply that the moderate risk is a 'transition' class which is assigned to users that do not belong to either 'low risk' or 'likes risk'.



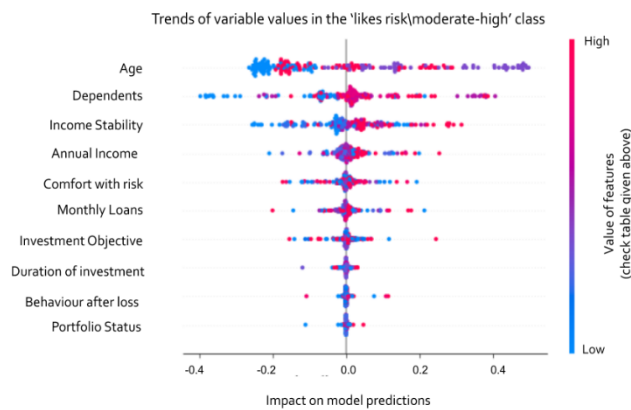
Moderate-Low Risk

Young users are less likely to belong in this class, while older users are more likely to belong. The mixed scatter of dependents and income stability show that these variables are not used for clear distinction. All other variables, 'red' values are less likely to belong in this class (users comfortable with risk, high income stability, who want to invest for 3-5 years with wealth creation goals). While, 'blue' values show that users tend towards tax saving goals, have high EMIs, low comfort taking risk etc. (refer to Figure 17).



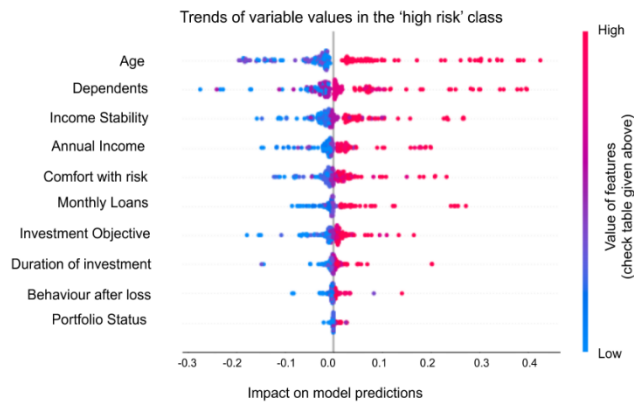
No Risk

A clear distinction between values chosen by users. Red and purple values show clear negative impact on this classification and blue values show positive impact. This implies that young users with low dependents and stable income most likely never belong to this class. Consequently, older users (55+), uncomfortable with risk, many dependents and low-income stability would probably benefit from being in the low risk class.



Moderately High Risk

In this class, being middle aged (purple and 35-55) has a greater positive impact on classification than being young. Further, mixed values of red and purple show more positive impact overall while blue and even red in some cases have negative impacts. This implies that the decision boundary for this class favours generally middle-aged users with no dependents, high income stability and a higher range of annual income.



High Risk

Compared to moderate risk a very clear distinction is observed in this class. This class favours young users comfortable with risk and with no dependents, high income stability, high annual income, low EMIs with wealth creation goals. On the other hand, not many distinctions between blue and purple values exist, apart from them having a generally negative impact. Model looks for specific cases with all criteria as positive.

Table 3.1: Figures show general trends across classes for each feature. The position on the y-axis is determined by the feature and on the x-axis by its average marginal effect on class decision. This is also known as Shapley values. The points represent samples taken from data. The colour represents the values of the features from low risk (blue) to high-risk impact (red). Please refer to figure 13 for different colours assigned to different variable values. Overlapping points are jittered in y-axis direction, so we get a sense of the distribution of the Shapley values per feature. The features are ordered according to their importance.

Mutual Fund details are conveyed similarly. The mutual fund data has fewer categorical variables compared to user risk as most data is numerical. For numerical data, the blue to red progression represents low to high magnitude values. These risk trend diagrams only show the features most important for class distinction. For detailed explanations of variables used refer to Appendix: Mutual Fund Quantifiers and Types.

Categorical/ Text based Variables	Values
Categories	<div style="display: flex; justify-content: space-between;"> <div style="width: 45%;"> <ul style="list-style-type: none"> Balanced Debt-FMP Debt-Income Debt-Ultra Short Term Equity-Arbitrage </div> <div style="width: 45%;"> <ul style="list-style-type: none"> Equity-Diversified Equity-Global Funds Equity-Sector Gilt-Long Term Gilt-Short Term Liquid </div> </div>
Primary Investment	<div style="display: flex; justify-content: space-between; align-items: center;"> <div style="width: 30%; border-bottom: 1px solid black; position: relative;"> Debt Money Market </div> <div style="width: 40%; text-align: center;">Equity</div> </div>
Investment Objective	<div style="display: flex; justify-content: space-around; align-items: center;"> Capital Wealth Income Tax Retirement </div>

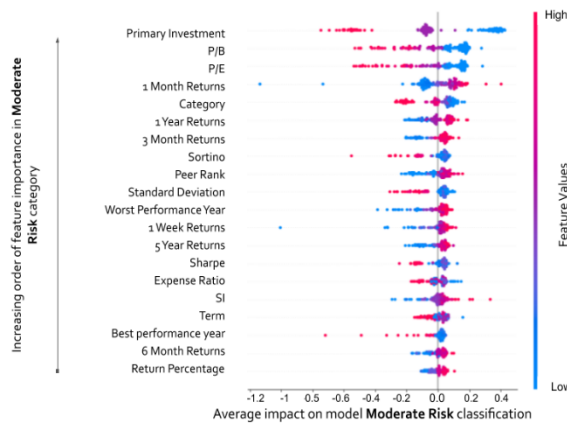
Figure 3. 14 Values assigned to variables

In Figure 3.14, the colours depicting values assigned to the categorical variables is shown. All other variables are numerical. This figure provides context for the mutual fund risk trends shown in table 3.2. The colour represents the values of the features from low risk (blue) to high-risk impact (red). Please refer to figure 17 to understand the significance and interpretation of colours for the model.

Table 3. 2 Trends captured by model during mutual fund classification

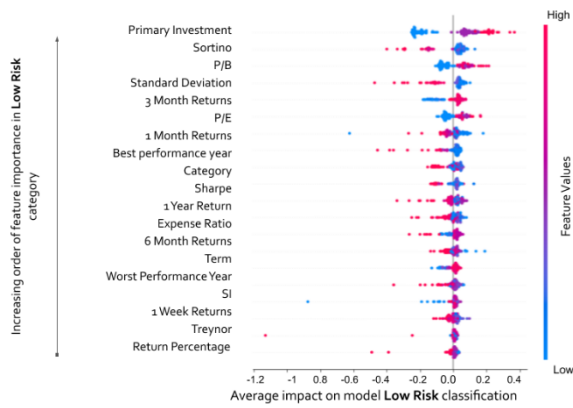
Trends in Mutual Fund Risk

Description



Moderate Risk

Primary investment is the most important decision factor, debt investment portfolios in categories of debt FMP, debt income/ ultra-short term, and balanced with low P/E and P/B ratios are favoured by this class. These funds typically have high returns, weekly and yearly. Although these funds have low Sortino, that is a low return on risk, these funds also have low expense ratios.



Moderate-Low Risk

These funds contain a mixture of many debt, money market, and equity funds with low Expense and Sharpe ratios i.e. the traditionally calculated return on risk is low. These funds seem to have mixed returns showing short term periods of lows and long-term highs. Additionally, low Sortino ratio indicates low risk adjusted returns which could indicate low risk in general. The high P/E and P/B ratios indicate that the stocks are expensive but have low risk steady returns.



No Risk

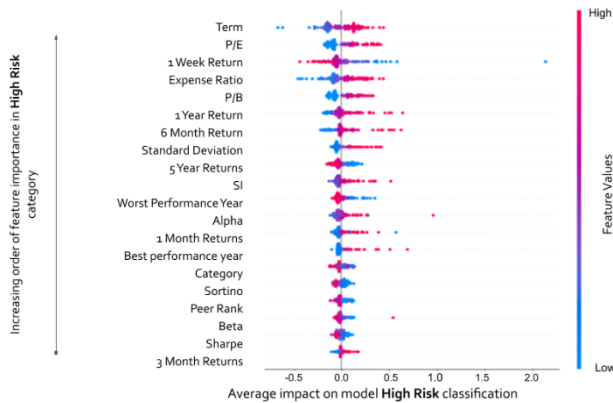
Short 'Term' gilt or liquid funds belong to this category. These stocks generally contain investments in gold. They do not possess much risk, but long-term returns are poor. High Sharpe and Sortino indicate good returns on any risk taken. Primary investment is in the money market. Low P/E, expense, and P/B ratios of funds in this class indicate that the stock has good valuation.



Moderately High Risk

Reveals some surprising trends, an overall low returns trend can be observed. This class generally contains debt funds with expensive investments (high P/E, P/B and expense ratios). The Sharpe and Sortino indicate a low return on the risk taken. All in all, if users

appropriately study this graph they would probably opt out of funds in this category. This demonstrates the need for transparent explanations.



High Risk

The most important distinction is done using the 'Term' of investment, the algorithm has learnt high risk funds encourage longer terms for adequate returns on investment. But one can observe, the returns are high for 6-13 months and then taper off. Therefore, matching duration of investment is important. The high P/E and P/B ratios indicate the valuation of these funds are relatively expensive.

Table 3.2 illustrates the trends captured by the model during mutual fund classification. This table also shows why users are recommended mutual funds based on risk, investment objectives and duration. One can see that apart from the most influential factors of the funds: category and risk ratios which are made to match user risk appetites and goals, the terms suggested for investment are also very important to guarantee reasonable returns.

These explanations convey complex intricacies captured by the algorithm. Therefore, they are technical and information heavy. While a few users might feel fatigued while following these graphs, they are necessary for complete transparency. As the goal behind the white box explanation approach works is complete availability and transparency.

4.2.6 Decision Boundaries

Decision boundaries are used by models to partition one class from another in the data and generally expressed as if-then rules/conditions. Explanations show users a decision tree based on training data and the predicted outcomes. The tree represents hierarchical rules followed by the model to detect and define categories. Users can interpret the decision-making logic followed by the

algorithm. For instance, in the diagram of our system below, if the user age is 45 with 3 dependents, the probability of 'low risk' classification increases. This method could extend to any algorithm.

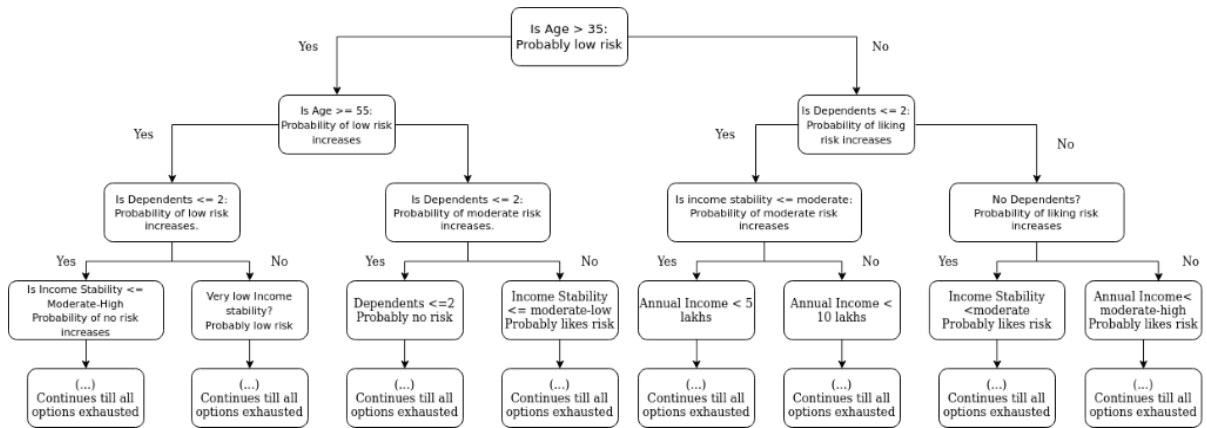


Figure 3.15 Top decision boundaries shown to users

Figure 3.15 shows the top decision boundaries shown to users. The flowchart illustrates the first few divisions made by the algorithm based on user answers. To measure user understanding in the usability study, users are asked to follow the chart for a sample user.

4.2.7 Group Comparison

Data contains groups or subsets such as different cultural, demographic, or phenotypic groups (e.g., race, geographic location, sex) and intersectional groups (e.g., age and race, or sex and age). Comparison of the individual outcomes (e.g., female) with an established benchmark group (e.g., gender: male). This explanation provides a relative group view for 'fairness' (Pedro Saleiro, 2018). The concept can also be used for defining the group or class of an individual user and answer contrastive-counterfactual user questions; "what did my neighbour get?" etc.

For the robo advisor, the users are encouraged to fill the questionnaire and re-fill the questionnaire in the system. For instance, if the user is unsure of his investment objective ("tax saving" or "retirement planning") the user can try selecting both to generate different profiles before finalizing. Observing the

effects of changing answers on risk and fund recommendations helps satisfy users' relative understanding of groups in the data.

White Box versus Black Box

Typically, as complexity increases, AI algorithms become black boxes and cannot be interpreted by humans. This increase in complexity due to an increase in volume, transformations or dimensionality, makes it impossible to comprehend the geometry of feature transformation or structure where thousands of neurons in deep learning algorithms work together for a solution (Bathee, 2018). Making white box interpretations of algorithms impossible. We explore the effect of such a case by constructing a black box alternative to our system equipped with the same strategy of explanations. The black box system uses more complex models with performance benefits equivalent to the white box system models.

Black box interpretability can be achieved through a technique called model surrogacy. Surrogate models are interpretable models trained on the original input data of a black box algorithm and the output generated by the black box algorithm on this data. These surrogate models have the ability to convey complex algorithms with accuracy and simplicity. This method of explanation also benefits companies who wish to protect their algorithm confidentiality. A fraction of data as a representative sample could be provided to a regulator to check the algorithm function (Krishnan, et al., 2020). Otherwise, the company could itself show users how their complex algorithms function through surrogacy models.

The key difference between white and black box explanations arises from the use of surrogate models. These explanations are speculative since a secondary model is tasked to explain the complex models' decisions, this leads to the danger that any explanation method for a black box model can be an inaccurate representation of the original model in parts of the feature space (Rudin, 2019). The difference between the explanations of the two systems is covered below. While it's effect on users is analysed through the survey.

Black Box

In our black box version of the system, a fraction of the data from classifications done by user risk and mutual fund models are utilized to build surrogate models. For this case we use polynomial SVMs for mutual fund classification and user risk calculation. For details on a comparative study conducted for algorithm selection, please refer to (Krishnan, et al., 2020). Black box explanations of SVMs are generated using an interpretable decision tree and random forest algorithm (refer to Appendix: Algorithm Performance Details). Due to the differing dimensionality and size of data, surrogates use twenty percent of user risk, and 50% of mutual fund risk data for surrogate model explanations.

Model Function	Assigning a risk category to users based on their preferences. This algorithm uses data gathered from past users, to classify new users into optimal risk categories according to their selected preferences.
Model Output	One of 5 risk categories: No risk, Low risk, Moderate risk, Likes risk, High risk
Number of Variables/Features used	10
Sensitive Variables	Age could be considered a sensitive variables.
Performance Metrics	Average Recall : 96%, Average Precision : 96%, Accuracy : 97%
Sources	Dataset and variables gathered from PayTM Money, Tavaga, EZMoney, Betterment, Wealthfront.
Sampling Strategy	Stratified sampling: proportionally equal sampling of data

Figure 3. 16 Black box version: Model details for user risk

Figure 3.16 shows the model details for user risk in the black box version of explanations. The requirements are satisfied, by listing algorithm function, performance metrics, sampling and expected outcomes. Additionally, sources of data have been mentioned. But details of the dataset are not revealed.

Model Function	Assigning a risk category to mutual funds to match and recommend users. This algorithm uses data gathered for a few mutual funds, to classify all mutual funds into optimal risk categories according to variables associated with each fund.
Model Output	One of 5 risk categories: No risk, Low risk, Moderate risk, Likes risk, High risk
Number of Variables/Features used	26
Sensitive Variables	No sensitive variables used.
Performance Metrics	Average Recall : 88%, Average Precision : 80%, Accuracy : 90%
Sources	Dataset and variables gathered from AMFI India, Money Control, Value Research, Mutual fund india.
Sampling Strategy	Stratified sampling: proportionally equal sampling of data

Figure 3. 17 Black box version: Model details for mutual fund risk determination

Figure 3.17 shows the model details for mutual fund risk determination in the black box version of explanations. A general idea of features used are given but the details of these features and working of the algorithm is not provided.

4.3.1 Feature Importance through Surrogate Models in Black Box Explanations

Surrogate models will not be able to capture intricate details like transparent or interpretable white box explanations. Since a fraction of data labelled by the complex model is used, only big abstract ideas will be correctly identified. We clearly state such approximations and limitations in the system.

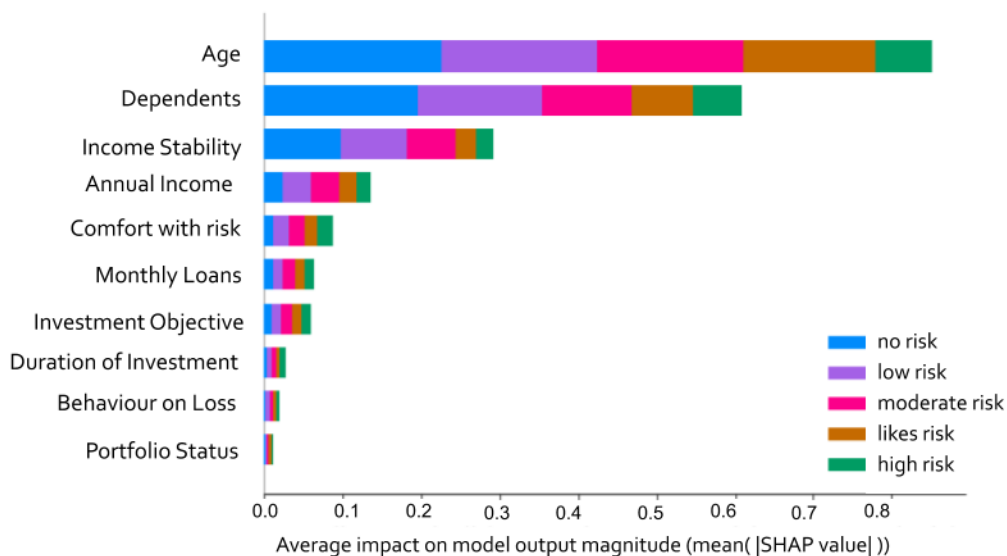


Figure 3. 18 Black box: Impact of variables on user risk category assignment

In Figure 3.18, feature importance of user risk is shown, as identified by the surrogate model. Although it was provided only 40% of original data (refer to Appendix: Algorithm Performance Details) due to the low dimensionality and balanced stratified sample provided for to this surrogate model the importance identified by the surrogate is precise and accurate to the decimal of magnitude. The order of importance and contribution of each variable towards each class is captured perfectly. Since this model uses fewer features the order of importance is captured accurately, while the impact of each feature towards each class is not calculated with the same precision as the white box explanations. As the feature importance decreases (towards duration, behaviour and status) the impact of each feature is difficult to analyse.

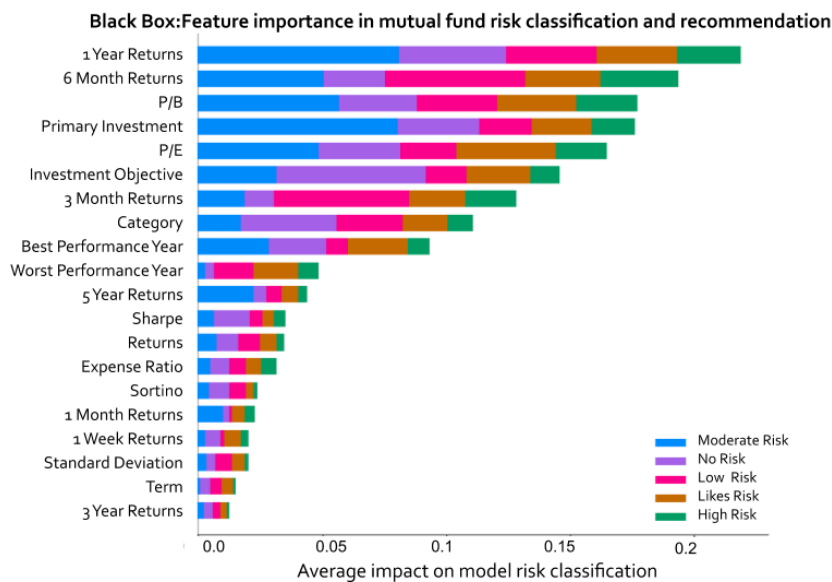


Figure 3. 19 Black box: Feature importance in mutual fund risk classification and recommendation

In Figure 3.19, feature importance of mutual fund risk is shown, as identified by the surrogate model. Due to the original sample size being low, the surrogate model was provided 70% of the data for examination. Even so, the surrogate has not been able to capture precise details of importance. When compared with figure 15, the order of importance and magnitude are off by a few values. Further, not all variables have been included such as the ones not showing much variation

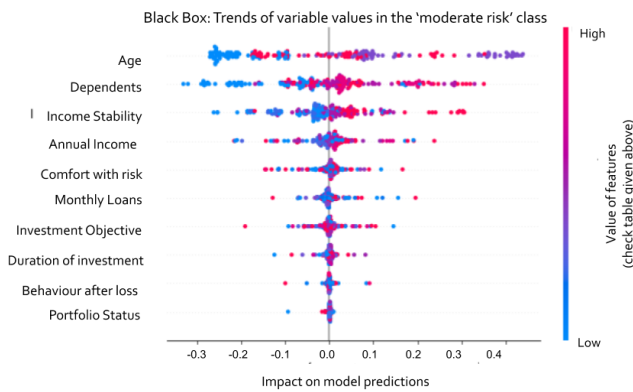
have been excluded. Such limitations of using a surrogate model should be clearly conveyed to the users. On the other hand, many trends have been captured correctly and mirror their white box counterparts. For instance, the surrogate conveys its observation of trends in the data, the 'moderate risk' class is the first check for all inputs, and area of primary investment, fund returns of 1 year and 6 months are the most important in determining this classification. Category variable is very important to the 'no risk' class.

4.3.2 Feature Effects and Behaviour

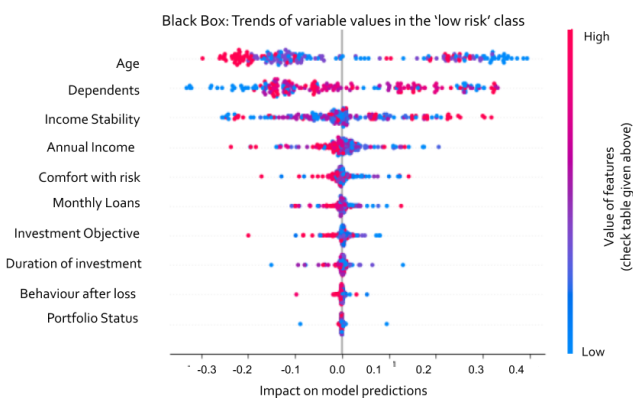
Feature effects are captured by the surrogate model to understand the global trends followed in data. Due to the increase in abstraction followed in surrogate model approach, approximate representations of these trends are available. They manage to capture the main trends accurately. The following tables highlight trends captured correctly while analysing the difference between their white box alternatives. The table also contains a comparison between these black box explanations and their white box counterparts. Although low dimensionality and sufficient samples, has allowed the surrogate to capture most prominent insights of the user risk model accurately, it proves to be more erroneous for mutual fund risk. An increase in variables in addition to smaller sample size causes the scatter to be mixed. That means distinct clusters of data that can be observed in the white box explanations are not as apparent in the black box samples. We predict that this would affect user comprehension. For both mutual funds and user risk classification, the colours depict the same information as shown in Figures 3.13 and 3.14.

Table 3. 3 Feature effects and behaviours captured by surrogate model of the user risk data

Trends in User Risk Classification: Black Box	Description
	<u>Moderate Risk</u>
	The interpretation of colour, order and jitter is the same as white box figures. However, the



distinction of values is not as clear as the white box. Although the trends are comprehensible. Users aged 18-55 with 1 to 2 dependents are put in this category. Annual income and its stability are generally high to moderately high. All details from these features are captured accurately. Not captured is the comfort with risk and investment objectives. Overall, the information conveyed is useful to get an idea of what belongs in this class.



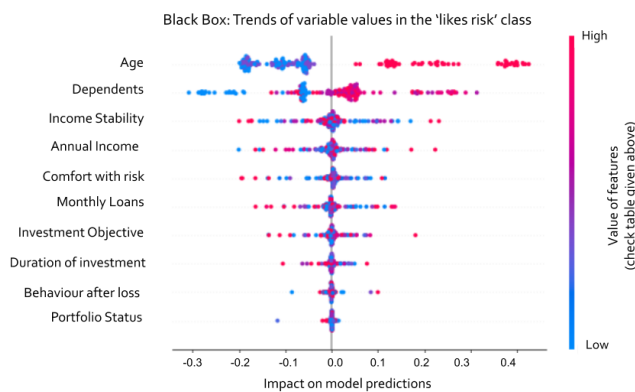
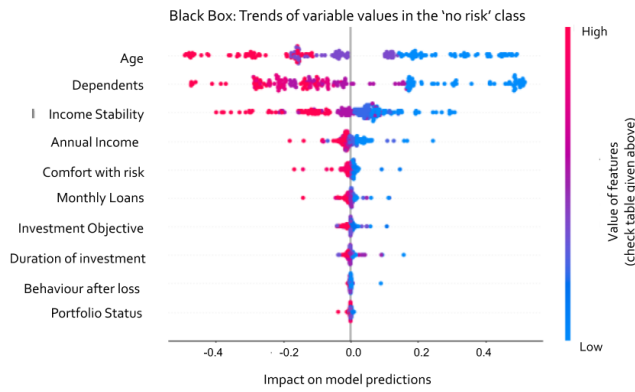
Moderate-Low Risk

Surrogate captures that older users with a higher number of dependents and lower annual income and stability belong to this category. Further, a low comfort with risk but low monthly loans as well. It is difficult to make acute distinctions without observing each variable closely. Similar to the graph above, the distinction of variable value and impact through colour is not as well

defined as its white box counterpart.

No Risk

Distinct values show, older users with 2-4 dependents belong to this class. Their annual income and stability are on the lower side. Further, they express low comfort with risk. Although all the variables that follow do not possess large distinctions, one can comprehend the nature and behaviour captured in this class.



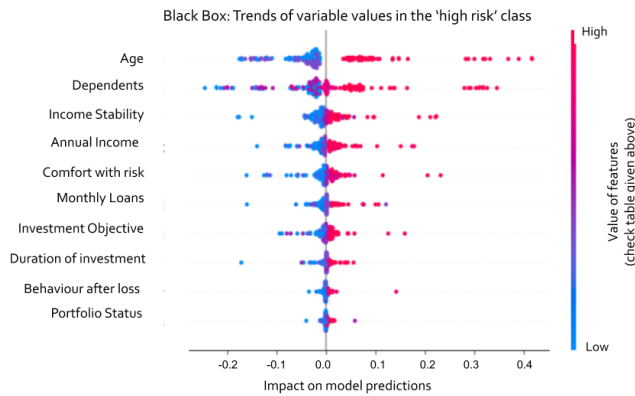
Moderately High Risk

Trends have observable shifts from other classes (moderate risk reference). Young users with no dependents belong to this class. They are not classified as high risk as all other variables of importance such as stability and annual income are mixed. Variables that follow stability are not

separated, as in the white box explanations.

High Risk

Distinctions for this category most closely resemble their white box counterparts. Young users with no dependents, high income, and stability belong to this category (red). They are comfortable with risk, have low monthly loans, and clear objectives of wealth creation or income generation. While as older users with low incomes and stability have a negative impact on the class decision. This means if a user is young and has more dependents the probability of belonging to high risk class decreases. The details of this class are accurately captured.



In Table 3.3, explanations show trends captured in the original input and white box algorithm output. They are compared with the original outcomes and white box explanations. Due to low dimensionality and enough samples, the surrogate has been able to capture most key insights of the model accurately. Y axis shows feature importance with a coloured dot representing a sample instance. The

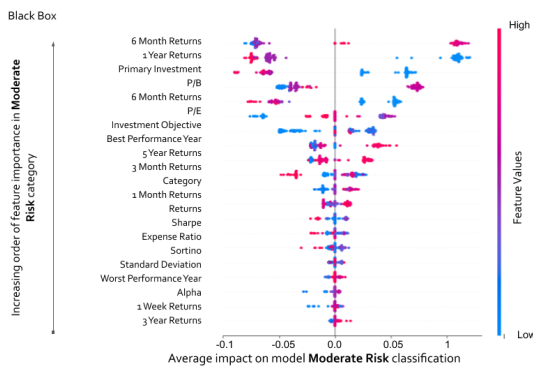
magnitude of positive effect of variable values in each class is shown on the x-axis and the jitter shows where instances lie on average. The difference between these explanations and their white box counterparts have been covered with each description.

Since the divisions of values in their contributions to each class are not strictly separated, it will be interesting to analyse the usability study to observe the difference in confidence and clarity felt by the users. The next table depicts surrogate model trends captured in the mutual fund risk algorithm.

Table 3. 4 Risk trends captured by the surrogate model

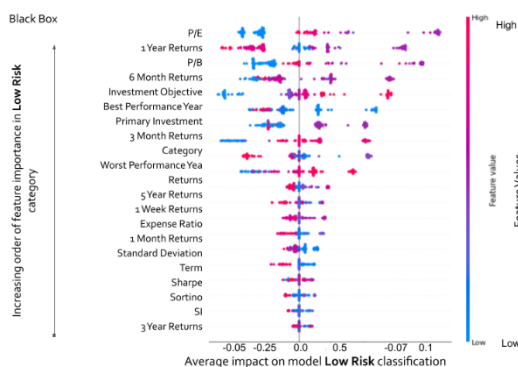
Trends in Mutual Fund Risk: Black Box

Description



Moderate Risk

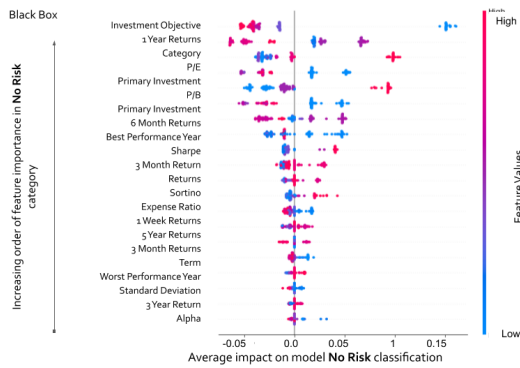
This class contains debt investment portfolios in categories of debt FMP, debt income/ ultra-short-term, and balanced with moderate P/E and P/B ratios. High returns are captured monthly and yearly. Although low Sortino and Sharpe that is a low return on risk values are not identifiable.



Moderate-Low Risk

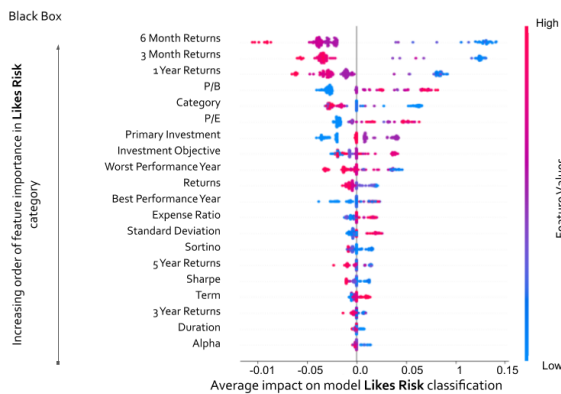
Funds captured are a mixture of many debt, money market, and equity funds with low Expense and Sharpe ratios. The traditionally calculated return on risk is low. These funds have

moderate returns instead of showing short term periods of lows and long-term highs (unlike white box). The P/E and P/B ratios indicate that the funds are moderately expensive.



No Risk

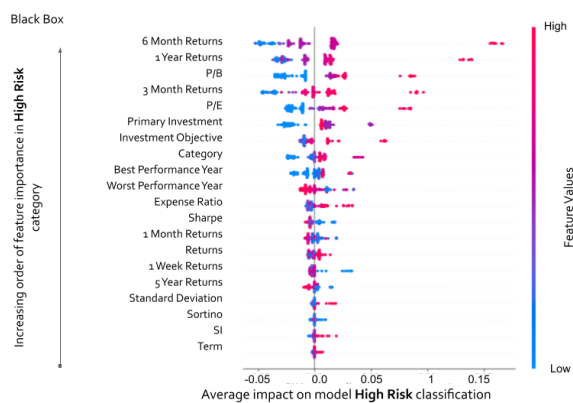
Correctly captured participation of short 'Term' gilt or liquid funds in this class. Further, high Sharpe and Sortino indicate a good return on any risk taken. Primary investment is in the money market. Low P/E, expense, and P/B ratios of funds in this class indicate that the stock has good valuation. These stocks generally contain investments in gold. They possess not much risk, but long-term returns are poor.



Moderately High Risk

The 'interesting' trends also observed in the white box version of this graph is captured here. The funds here are expensive with high expense, P/B and P/E.

Also note low yearly and monthly returns coupled with lower return on any risk taken. Category shows these funds contain balanced, debt and a few equity funds (blue and purple). The surrogate model could capture the limitations of this fund accurately.



High Risk

Distinctly observable trends have been captured for the 'high risk' class. For instance, the trends across high returns for 1 year, 6-3 months, and return on investment are correctly captured. Further, accurately captured are the primary investment of these funds is in liquid to gilt in the money market categories. Low Soritno and Sharpe (blue) indicate a lower return on risk. Term of investment is not recognised as the deciding factor but the gist of requiring higher term (red) investments is captured. Overall an accurate representation of data is made, although the estimation of original

model logic is approximate.

Table 3.4 shows the risk trends captured by the surrogate model. Due to low sample size and high dimensionality the surrogate model provides correct representation of data but an abstract idea of the original model logic. Although, it is captured precisely and accurately for classes on the extremes such as no risk and high risk, the surrogate seems to have trouble deciphering intricacies of the variable contributions of the middling classes.

These black box explanations succeed in conveying **approximate** feature importunacies and trends in the data. The goal of using two systems for user study is to analyse the difference in user experience and comprehension between the transparent white box and surrogate black box approximations.

Experiment Design of User Study

The broad aim of this user study is to gather data on the needs of users of complex algorithmic systems, then further translating them to generalised user requirements that support the development of useful and usable explanations. Conducting such research on users is necessary to build user-centric systems. The robo advisory system and its explanations should be inherently user-centric, therefore it is essential to evaluate whether the user understands the logic conveyed through these explanations. To recap, data on three key questions are garnered through the user study: (i) whether users can discern the entire system along with its advantages and limitations, (ii) if there is a difference between white and black box explanations when it comes to comprehension and usability, and (iii) whether this comprehension positively impacts the usability of the system. While the previous section explains the setup of experiment task, this section conveys the design of the experiment. First, we analyse the trends seen across similar experiments. Second, we design the qualitative and quantitative survey that is used to analyse user comprehension, explanation usability, system usability and trust. Third, we explain the sampling strategy used to select

participants for the study. Participants were recruited through a voluntary participation online survey shared through social media. User research in XAI is common practice, a number of studies consulted during the design of the experiment are covered in the Appendix: Similar User Studies on XAI section. These papers highlight the value of empirical application specific investigations and human evaluations of XAI. The standard size of participants used in each study is 15 to 20. The participants are primarily involved to gauge the usability of the system and its explanations.

Similar User Studies on XAI

User research in XAI is common practice. This section covers many studies consulted while designing our user study. These papers highlight the value of empirical application specific investigations and human evaluations of XAI. Usability (Bekker, 2000) is a key concept in understanding the extent to which a product (here, the explanations) can be used by the target audience of robo advisory systems with effectiveness, efficiency, and satisfaction (ISO 9241-11: 1998; ISO 13407: 1999).

(Danding Wang, 2019) implements an explainable clinical diagnostic tool for intensive care. Designed for physician usability where the ease of use was determined through user testing, the user study conducted with 14 physicians shows a high interest in knowing the importance of features used by the algorithm as well as understanding the answers to 'what if' questions satisfied by counterfactual reasoning. The study found that physicians repeatedly seek out raw data to verify explanations, also found that users formed a hypothesis and verified it with provided explanations. This insight supports the decision of, including thorough instructions, on how to read explanations of the robo advisory system. Users can come to their conclusions of the system logic after they comprehend the explanations.

Similarly, Intellingo adopts a user-centric approach towards designing explanations for intelligent translation software. Their user study considers 26 professional translators and experts and reveals that displaying more information

to enhance intelligibility had a positive influence on user experience (Sven Coppers, 2018). Another example is (Malin Eiband, 2018), an AI-based personal fitness coach enhanced with the ability to provide the rationale behind its recommendations. The interactive application aims to aid product usability and comprehension by constructing better user and expert mental models, done by presenting guiding questions to users. To determine how and what to explain, an iterative study was done with a team of 8 experts. Many similar such user studies were conducted by XAI researchers DXPLAIN (G. Octo Barnett, 1987), LIME (Marco Tulio Ribeiro, 2016), etc.

Details on the survey design

Sampling Strategy

The evaluation of robo advisory explanations would ideally require a sample that mimics the real target audience of such applications. We use sampling strategies from purposive sampling. Based on a review of strategies selected by consulting wealth advisors, a robo financial advisory firm and reviewing similar literature. The iterative model of participant selection is similar to *theoretical sampling*. Theoretical sampling is different from many other sampling methods in a way that rather than being representative of population or testing hypotheses, theoretical sampling is aimed at generating and developing theoretical data (Glaser, 2012). Through continuous and iterative reviews of the data we collect from users, we come up with plausible hypothesis and implications. The study is not designed with a singular hypothesis in mind. Through our first iteration of users, we gather key concepts and areas of division that could lead to relevant and impactful results.

In the *quota sampling* strategy to select our first batch of participants. We choose individuals according to a wide range of traits and qualities such as: differing age groups, dependents, incomes etc. This sample population consists of beginner to advance investors with varying ranges of age (25 to 70+), dependents, and income. The sample puts an emphasis on Mumbai, Bangalore, Delhi, and Pune,

as these cities account for the majority of stock investments in India. This is a commonly used approach in user research and can be found across many similar studies (refer to Appendix: Similar User Studies on XAI). The study begins with six iterations of five users. Based on the target population, users differ on two key issues, features (age, dependents, etc.) and experience (seasoned investors, novice/beginner investors). These divisions determine the sample population and iterations. The first three iterations depend on the primary feature of the system, age, and further selections of varying income and dependents. After analysing these answers and making any required changes, the next divisions depend on expertise or whether the system and explanations are usable for seasoned/mature investors as well as novice investors.

These iterations are followed by *snowball or referral-based sampling strategy*¹⁶. Through this strategy, existing users recruit future subjects through acquaintances, that is, participants from the first batches share this online survey with friends and family. This allows us to access a large population with mixed qualities. The same sampling strategies are used to evaluate both white and black box systems with two different sets of users.

For each batch, analysing results from about five users suffices when it comes to gathering relevant insights, beyond which the observations from further users become repetitive, as Nielsen (2000) noted. Since one of our main objectives is to focus on the reaction of different demographic groups to explanations, we gather and analyse the results for at least five users in each group.

Results

The results of the user study are divided into seven parts. First, we analyse the background of user study participants. Second, we evaluate user comprehension of explanations through the quantitative sections of the survey. Third, we measure usability of explanations as well as the robo-advisory system. This is

¹⁶ Snowball sampling is a nonprobability sampling approach in which current research participants recruit prospective study participants from their own social circles. As a result, the sample group is seen to expand like a snowball. As the sample grows, enough information is collected to be helpful for the study.

done using responses to a combination of qualitative and quantitative questions of the survey. Fourth, to study the influence of explanations on user psyche, we analyse the opinion gathering qualitative aspects of the survey. Fifth, we measure the effect of complexity on explanations, comprehension and usability. Black box system uses more complex algorithms for the same task, comparison with white box allow us to evaluate the effect of change. Sixth, difference in user comprehension and opinions of explanations based on different demographic backgrounds is evaluated. Seventh, we comment on the broader implications of this study. The contribution user comprehension, opinions and usability towards the broader picture of system explainability and innovative inclusion.

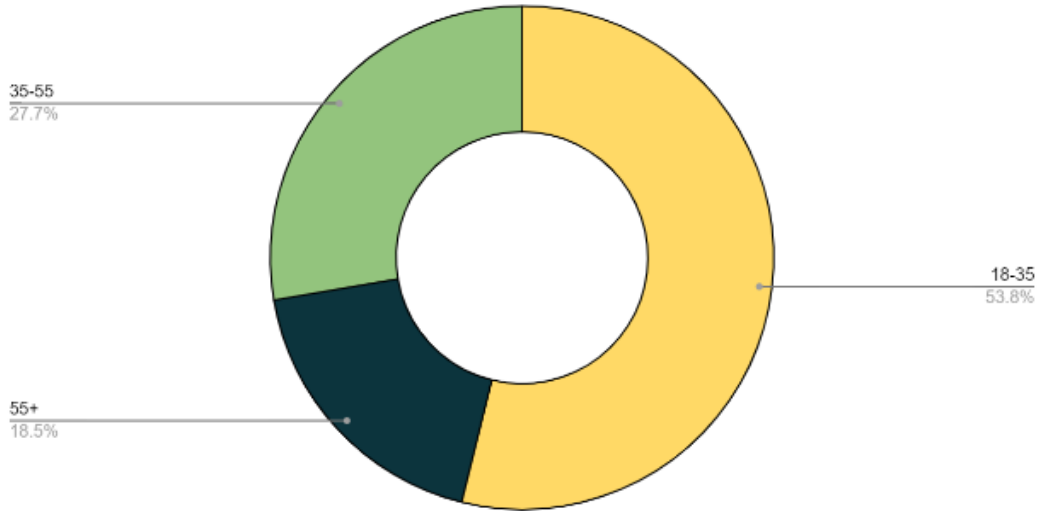
Participant demographics

From a total of 105 participants in the study, 54 users were shown the black box system and 51 the white box. Survey was conducted for a duration of four months, each participant received ample time for testing and analysis of explanations, similar to real mobile application users. An online survey with voluntary participation was shared through the authors' personal networks and social media platforms. The initial sample recruited co-workers and acquaintances that fit a criterion for sampling iteration (novice users, seasoned users etc.) (refer Sampling Strategy). The survey had participants from mixed nationalities, a majority from various cities in India and a few from Germany and United States.

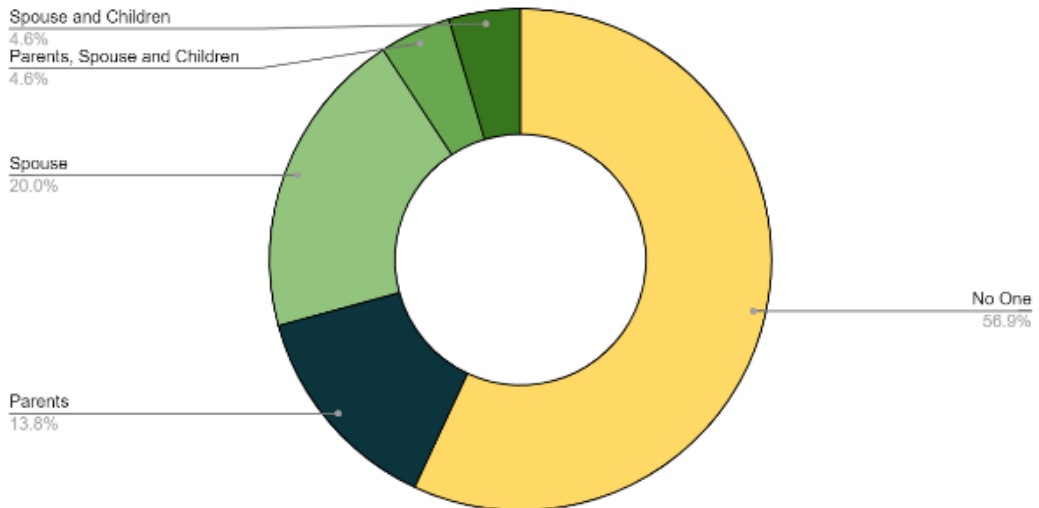
The primary characteristics important for user risk profiling in our robo-advisory system are: age, dependents and annual incomes. Please refer to the table on sampling strategy in the Appendix section for a detailed breakdown of the figures. In both white and black box studies, 51% of the users are 18-35, 30% are 35-55 and 20% over 55. Majority of participants were young. Out of all participants, 40% of had no dependents, 28% had 2, and 24% with a single dependent. Only 8% had 3 or more dependents. In general, a majority of participants reported having moderate (47%) to very low (23%) incomes, given a scale of 0 to 25 lakh INR. The results of both systems are comparable since the participant backgrounds and demographics are similar, barely varying by a few

percentage points. Detailed divisions and differences in the two systems is shown through Figure 3.25.

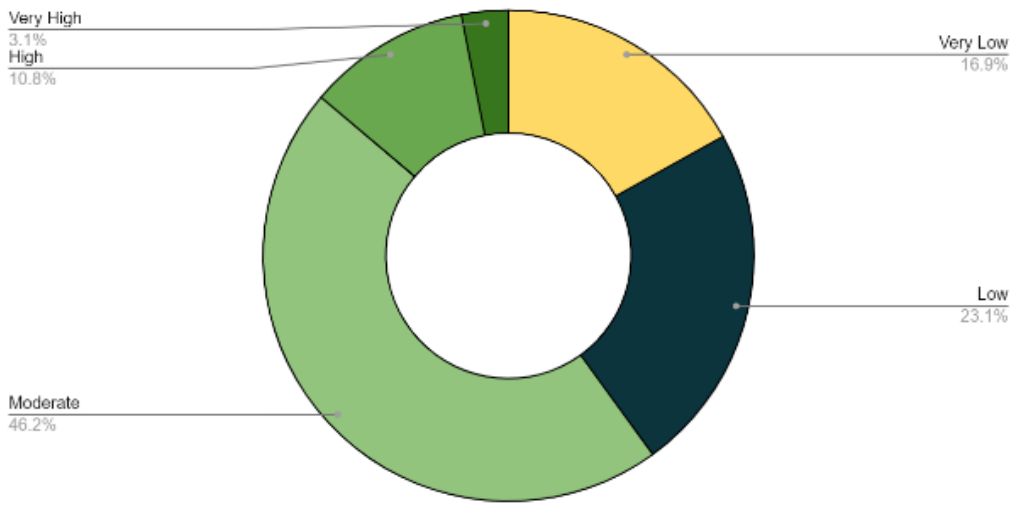
A: White Box User Age



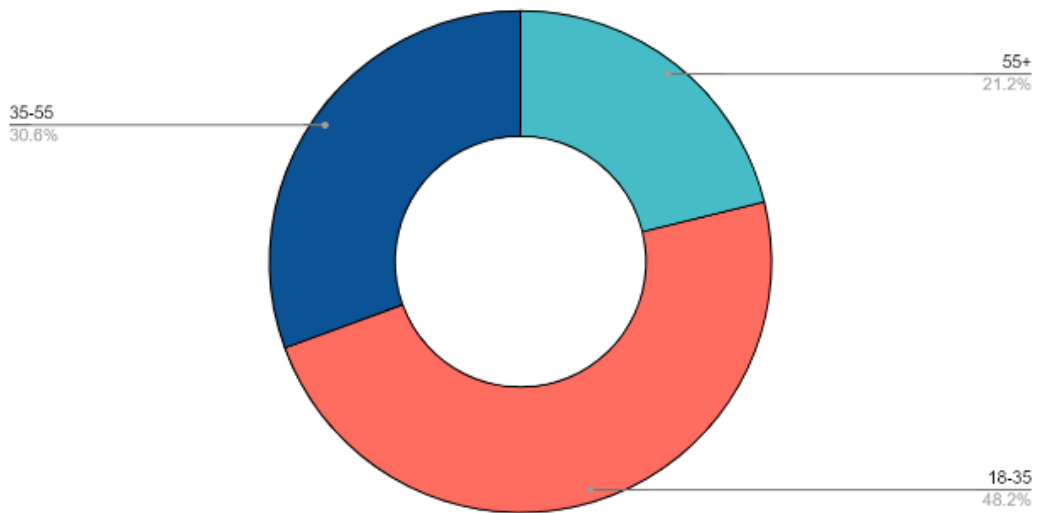
B: White Box User Dependents



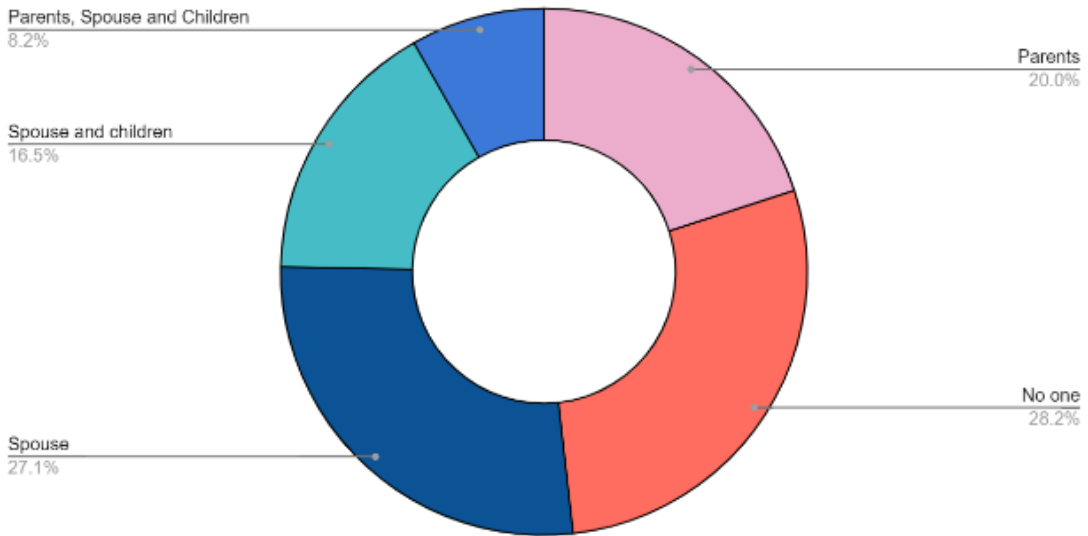
C: White Box User Annual Income



D: Black Box User Age



E: Black Box User Dependents



F: Black Box User Annual Income

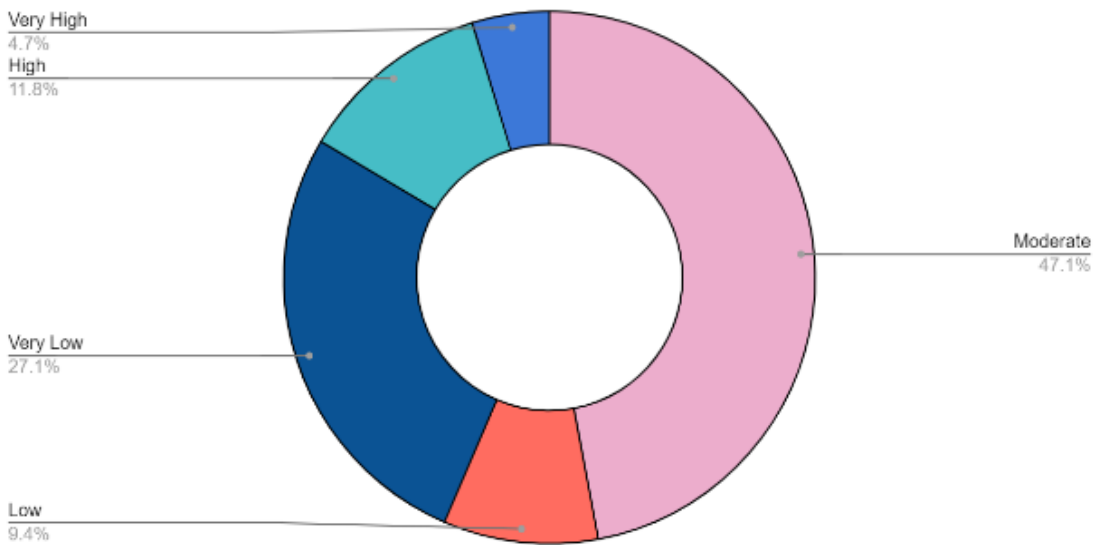


Figure 3. 20 Comparing users from black box and white box studies

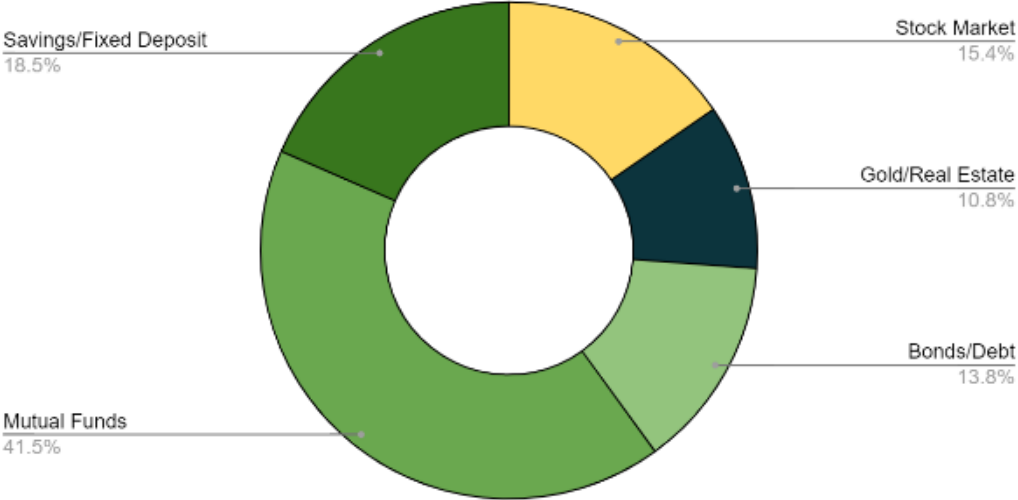
Figure 3.20 (A to F) compares users from the black and white box studies based on major demographic divisions. (Left) White box demographics show that the majority of users were 18-35 years of age, with no dependents and moderate income. This is the same majority showed in the black box demographics (Right).

Therefore, the results are credibly comparable since they consider users with similar backgrounds.

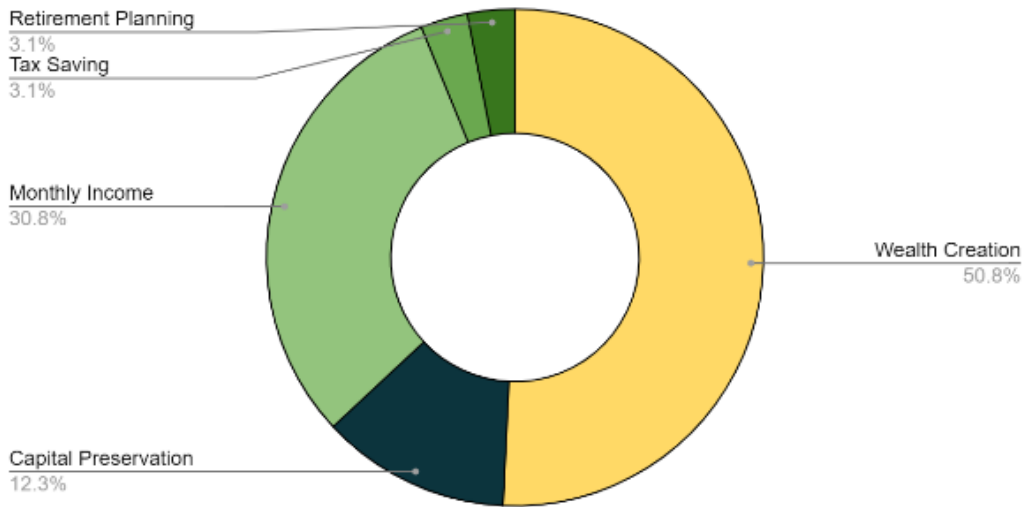
Secondary characteristics important for risk profiling and fund recommendations are current status of participant portfolios and future investment objectives. 43% of participants from both systems have current investments in mutual funds. Followed by 18% relying on fixed deposits or savings and 6% on gold or real estate. This implies a majority of people are currently invested in low risk funds. Only 17% of participants have higher risk investments such as stocks or bonds. In contrast, a large number of participants want to invest for the objective of wealth creation (36%) or capital preservation (23%) or monthly incomes (32%). All of which require long term investments or large short-term investments in funds commonly classified as higher risk. However, only 36% of participants are willing to invest for more than 5 years and only 29% are willing to invest in high risk funds for less than 3 years.

This contrast between objectives and current investments can be observed further with a hypothetical question posed to participants. When asked what they would do after losing 20% of their investment in a year, 62% said they would sell immediately or after market stabilises while, 17% would keep their investments. Finally, when asked directly about their risk preference, only 9% of the participants reported being uncomfortable with risk. Fortunately, the system considers all participant answers and the contrast between objectives, current investments, and preferences is taken into account during risk profiling and recommendation.

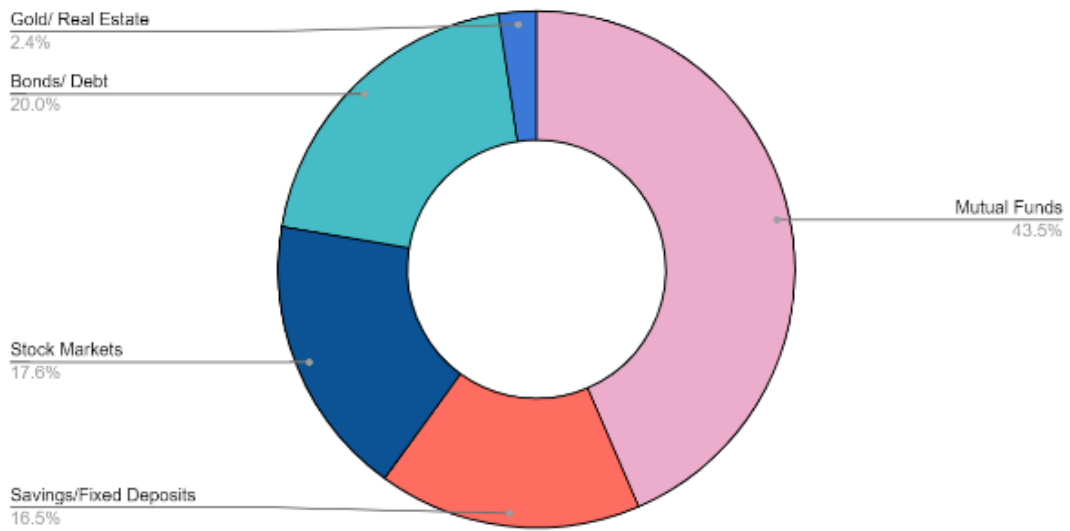
A: White Box User Current Portfolios



B: White Box User Investment Objectives



C: Black Box User Current Portfolios



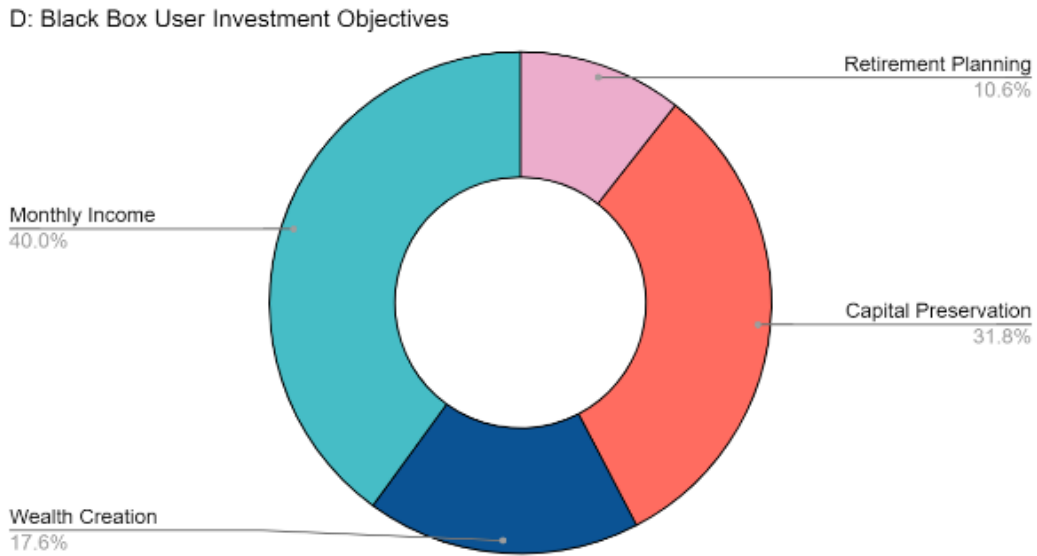


Figure 3. 21 Demographic divisions

Figure 3.21 (A to D) shows demographic divisions based on two of the most important features for robo advisor mutual fund recommendations. Investment Objectives and current portfolio status of users in the black box (Right) and white box (Left) studies. Both studies have majority of users investing in mutual funds with objectives of either monthly income, wealth creation or capital preservation. Therefore, overall, most participants are young, with few dependents, mostly stable incomes and a self-reported interest in wealth-based investments -- 52% are recommended moderate-high or high (in cases of high income and high stability) risk funds by the algorithm. Figure 3.22 depicts the majority characteristics of participant demographics.

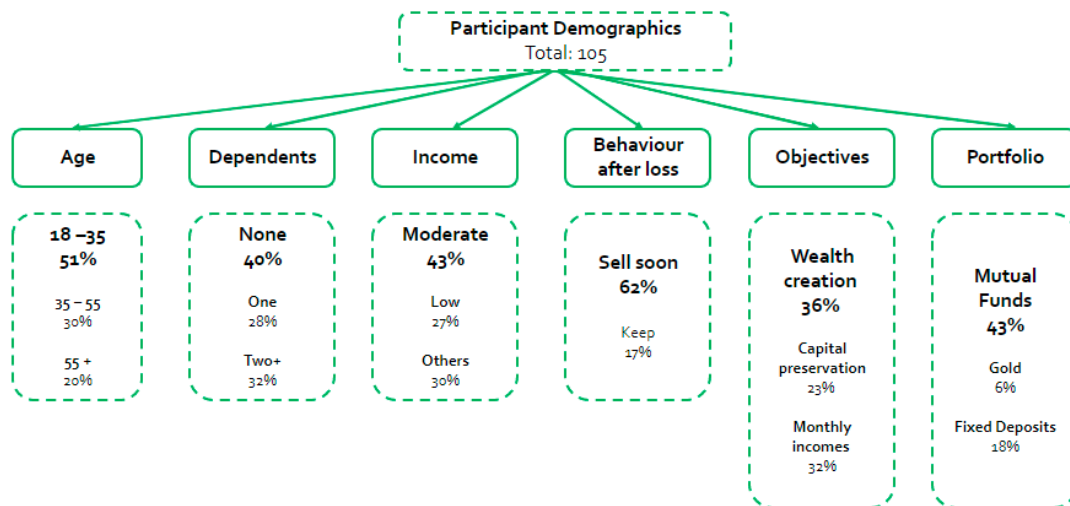


Figure 3. 22 Majority demographics of survey participants

User comprehension of explanations, recommendations

User comprehension is tested using the quantitative survey with close ended questions that require participants to analyse all explanations. Questions relevant to this section can be viewed in Table 8. This section lists our statistical and analytical findings, the implications of analysis are detailed in the “Usability” section.

The quantitative survey begins with localised explanations and recommendations. This required users to identify the most important variables for the risk profiling algorithm from a personalised bar plot (refer to white box explanation). For this, 66% of the white box participants answered correctly. The result for black box was similar, with 64% accurately answering questions on localised user risk. Immediately after this explanation, participants were given a textual description of important variables for mutual fund recommendation. Only 31% white box and 38% of the black box participants answered this question correctly (Observation 1).

Next, participants are shown model specifications for the user risk and mutual fund risk algorithms. These model specifications convey details such as performance metrics, function, sampling strategies etc. (refer model details and descriptions). 63.8% of the white box participants answered model specifications for user risk correctly, while 73% were correct for mutual fund specifications. For

black box, 51% were correct about user risk and 53% about mutual fund risk. There were minute differences between the specifications for two systems, except for the performance metrics of the two algorithms (refer White Box: Model Specifications and Black Box: Model Specifications). Therefore, the very small difference in comprehension could be explained by the fact that both the systems displayed performance metrics, sources and distributions of data, only differing in the extent of variables shown (Observation 2).

The global feature importance diagrams show the importance of each feature in each outcome class. Meant to provide a generalised-holistic view of the important feature in the system. This is an average generalised importance, differing from the personalised local importance diagram. The white box global feature importance met with 80% accuracy. While the black box users had an accuracy of 53%. In each of the systems, global importance of user risk and mutual fund risk is presented to participants. The percentage calculated is given as an average of two algorithms (user risk and mutual fund risk). The largest difference in comprehension is between mutual fund risk algorithm questions for the two systems. 73% of white box participants got them correct compared to 53% of black box. Mutual fund risk algorithm makes use of a large number of variables. Due to the approximation used by a surrogate model in the black box algorithm, these details are captured with a lower degree of precision. Therefore, lower participant comprehension is not due to an error in understanding the graph, it is due to the surrogate algorithm. (Observation 3)

The decision tree (Figure 3.19) displays abstract logic used by the algorithm in a simple human understandable format. This explains the causal logic and boundaries used by the model to make decisions. 77% of white and 83% of black box participants followed the logic correctly. These graphs follow linear logic explaining the high rate of comprehension. However, the decision trees showed in the system only covers five layers of depth used by the algorithm. As algorithms increase in complexity and variables decision trees cannot provide an accurate idea of the decision making. Therefore, this explanation can be used in only a few instances. (Observation 4)

The feature effect and behaviour diagrams are the most detailed and complex explanations shown to participants. These explanations allow an in-depth class-wise global comparative view of feature behaviour learnt by the model. Participant comprehension for white box was 67% for the user risk algorithm and 57% for the mutual fund risk algorithm. The higher number of variables used in the latter proved to be a source of confusion for the participants. The participants voted these explanations hardest to comprehend. However, despite the high technicality and complexity of these explanations, a large number of white box participants have answered accurately. (Observation 5)

Amongst the black box participants, 34% were accurate for user risk questions and 26% for the mutual risk questions. This is a drastic decrease compared to white box. The surrogate algorithm used in the black box system provides an approximation of logic. Therefore, intricate details are not precisely captured by the surrogate. This is reflected in the graphs through reduced distinction of variable scatter. This effect can be observed through a reduction in well-defined boundaries of the variable values, Tables 3.2 and 3.4 show the increase of mixed coloured jitter plots in the black box explanations. As the explanations go into deeper details the surrogate model of the black box system cannot decipher the transformations with the same precision as white box. This is due to the differences in transparency and complexity that are the result of a trade-off between precision and interpretability while using surrogate models. This is a significant finding because the reduction in precision has directly affected user preference. Although approximately accurate, this makes it difficult to follow the logic of the graphs. (Refer to white and black box "Feature effects and behaviours" section). (Observation 6)

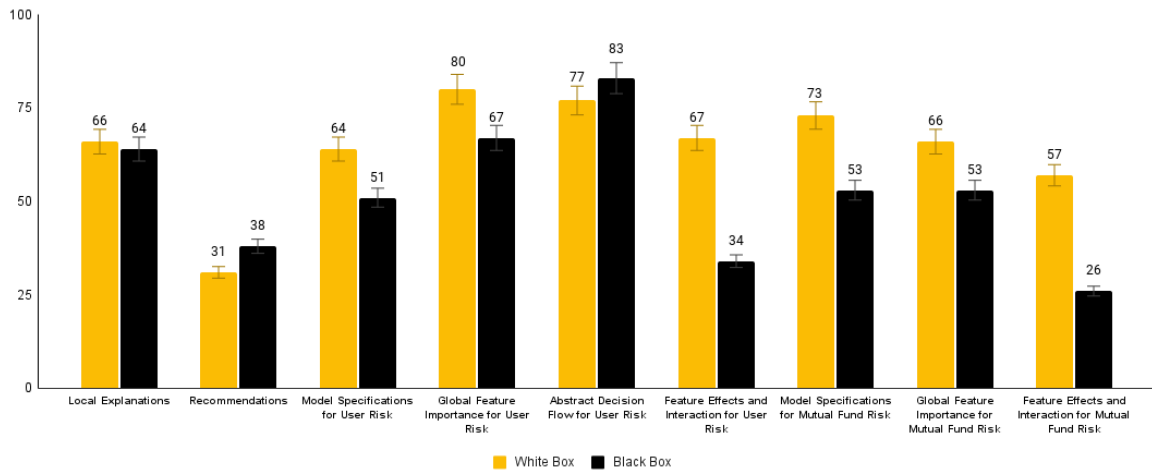


Figure 3. 23 Comparing black box and white box explanations for comprehension and accuracy

Figure 3.23 shows the difference in user comprehension and accuracy for white and black box explanations. This difference is statistically significant at 90% confidence level¹⁷, verified through T-test¹⁸, suggesting that white box explanations provide greater comprehension and accuracy when compared to black box explanations. The data is normalised to percentage values. For both systems, the 95% confidence interval lies between +/- 5 points.

Effects of explanations on participant opinions or perceptions

A part of the qualitative survey gathers the system users' opinions and thoughts on the explanations and the system. The primary goal of these questions is to understand participant psyche. This includes behavioural aspects (such as challenging areas, points of frustration, fatigue, etc.), general thoughts on specific explanation difficulties or usability, and measuring the effect of explanations on user trust. Questions related to user opinions on explanations are covered in "Usability of Explanations".

¹⁷ We have taken the liberty to use 90% confidence interval for the T-test as it is a relatively small sample size and because the field is still under-explored, making this dataset original.

¹⁸ $H_1: (White\ box - Black\ box) > 0$
 $\Rightarrow Pr(T > t) = 0.0615$

T-test results for white and black box explanations for comprehension and accuracy is available in the Appendix.

The first open ended qualitative question was asked after participants of the research were classified into a risk category. Most (70%) agreed with the first assigned risk category (Observation 7). The rest reported *feeling* unsure about being classified into a category of risk they perceived to be higher. However, participants were allowed to change objectives and durations of investment. The effects of which on the risk categories were shown through the local feature importance. Participants were encouraged to change their answers in addition to a question dedicated to guide them through the effects of experimenting with the risk determination survey. Surprisingly, not many people attempted to change their final answers to get into a more comfortable risk category. Only 10% of the users changed their answers out of curiosity, beyond the survey requirements (Observation 8).

The second round of questions were used to gauge user expectations from explanations. 68% of users felt that it was important for them to know how their selected preferences (or features) affect model decisions. (Observation 9).

Next, a set of questions at the end of the survey asked participants about their opinions on the presence of explanations in general. 90% of participants on average agreed that all algorithmic decision-making systems should come with explanations. 84% would like to see the explanation framework used in this paper across ADS for robo advisory applications. (Observation 10). A breakdown of these questions is shown in Figure 3.24.

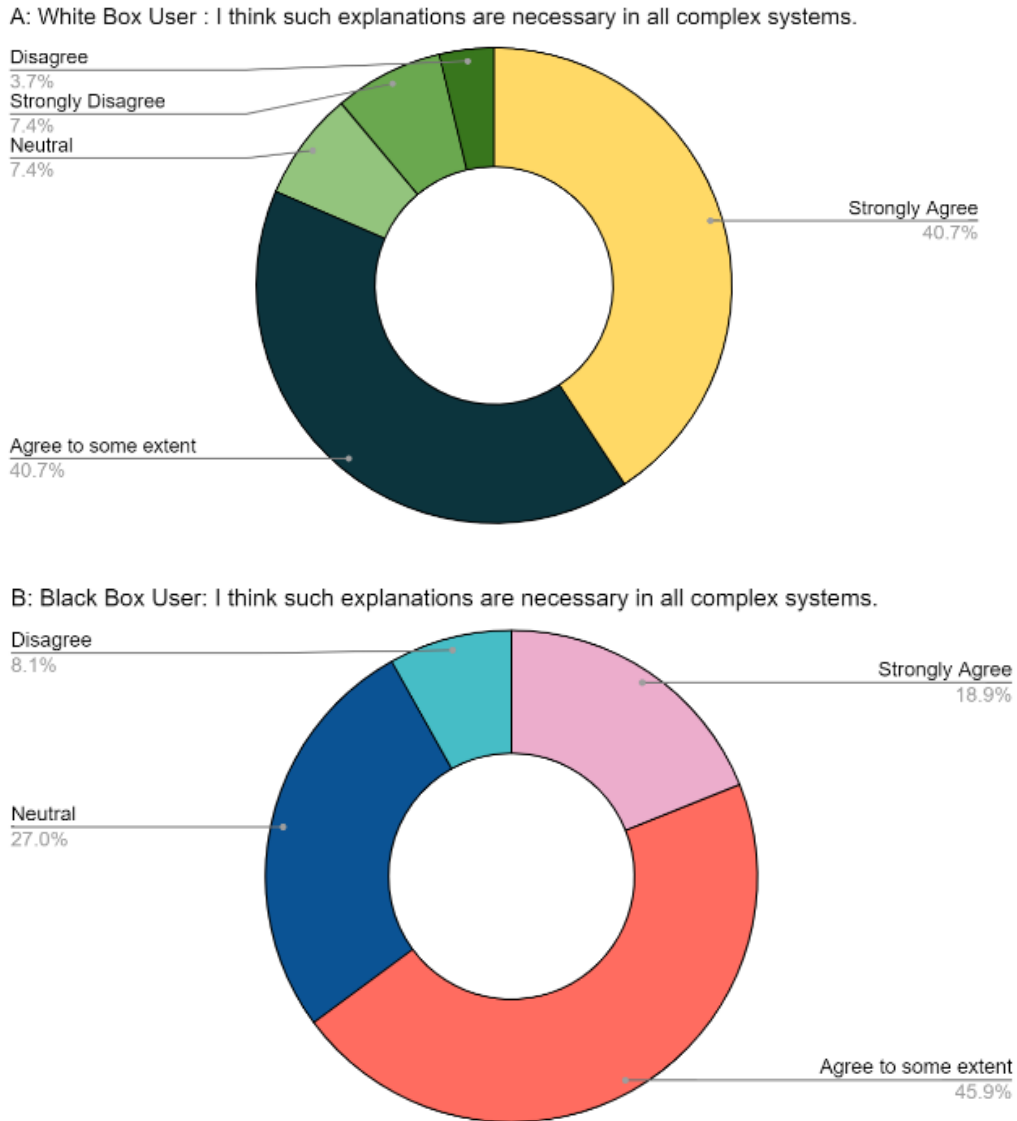


Figure 3. 24 Difference between user response in white box and black box scenarios

Figure 3.24 shows the split specifically in the answer, "I think such explanations are necessary in all complex systems," general opinions on explanations of AI. An interesting difference between the extent of agreement between white and black box users, shows a 20% increase in neutral positions and a 22% decrease in agreement.

Participants reached a point of frustration around the mutual fund risk explanations of feature behaviour and effects. Around 9% of white box and 11% of black box quit the survey at this point. This, in addition to observation 5 and 6 - the reduction in accuracy for detailed black box explanations, we suggest either

minimizing the complexity of feature behaviour or making these explanations optional for interested users, especially for black box explanations. (Observation 11)

Based on these results, especially observation 10 and 11, we propose that explanations would undoubtedly be a welcome addition to any decision-making algorithm that requires user information. A basic explanation relating the user information and its effect on the outcome would be the most popular and intuitive.

Usability of explanations

Usability of explanations determines whether the explanations are effective and interpretable. This is measured through a combination of results mentioned in this section. User comprehension determines the interpretability of explanations. User opinions and trust determine the effectiveness of explanations.

To determine the most popular explanations participants were directly asked to rank the helpfulness of explanations. Their options covered individual and combinations of the following explanations: feature importance, behaviour, effects, model specifications, and decision logic tree. 40% of black box and 29% of white box participants found all explanations useful. An almost equal number of white box participants prefer feature-based explanations. 28% explicitly prefer feature importance and interaction-based explanations. In stark contrast to only 8.5% of the black box participants. However, this can be explained as the results of approximation due to surrogate models and low comprehension of these explanations. 30% of the black box participants found model specifications most useful. Only 9% of white box users found these useful. User comprehension for white box feature interactions is amongst the highest (based on observations 3 and 5). Participants of both systems show relatively high accuracy in answering model specification questions (based on observation 2). However, user comprehension of certain feature effect diagrams is amongst the lowest accuracies throughout the systems (based on observation 6).

This dichotomy in preference is propagated further when users were asked to list the limitations of the system; most users (75%) reported reducing, changing or omitting parts of model specifications. Therefore, in another question specifically about model specifications, 20% did not mind using decision making systems with just model specifications. Out of these a large majority were black box participants (80%). Although many users found model specifications useful, this is a low number. Specifications are vital indicators of real-world performance. While these metrics have meaning for developers and regulators, they hold little to no meaning for users of the systems. On their own, model details provide an overview of the model performance but do not divulge a lot of information regarding the decision logic. They can only impart the user with a preliminary idea of the model, its intent, functions, and performance. Ultimately, user preferences seem to be highly dependent on the explanations themselves. In cases where there is a clear and distinct explanation it adds value to users. Otherwise, external data or features that are familiar to users should be shown to help users understand and trust the system (such as the data sources in model specifications).

The decision tree is the least useful explanation according to users. Only 3% of white box and 6% of black box users found this explanation useful. This is interesting since, observation 4 shows this explanation is easy to understand and user comprehension of decision tree logic was remarkably high in both systems.

Participants were asked to rate their understanding, before providing explanations and after. For white box, an average rating of 5.8/10 was given just after the local explanations and a rating of 7.7 given after showing all the explanations, showing an average increase of 2 points. For black box, the average increase in before and after is less than 1 point. (Observation 12). This reinforces our belief that explanations do have a positive effect on user opinion of the system.

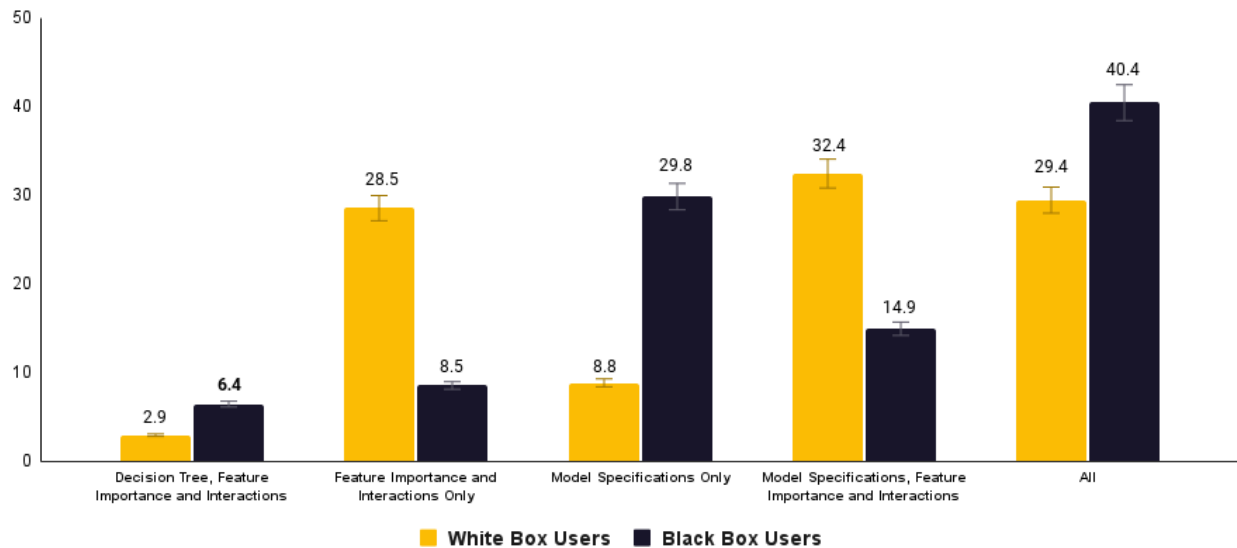


Figure 3. 25 Participant preference of explanations

Figure 3.25 shows participant preference of explanations. The chart shows that majority of users like all explanations. However, model specifications have a 21% increase in preference for black box explanations. While feature importance and interactions show a 18% reduction in popularity. The data is normalised to percentage values. For both systems, the 95% confidence interval lies between +/- 5 points.

Participants were asked to suggest improvements to the explanation framework. 23% of the white box participants and 35% of black box mentioned a need to simplify explanations. In this open-ended question, most participants explicitly mentioned their difficulty in comprehending feature effect and behaviour type explanations. While this is an understandable suggestion since this type of explanation is highly technical, most users seem to have correctly comprehended the data shown (based on observations 5 and 6).

Another improvement to the framework is to reduce the number of explanations provided and the time required to understand these explanations. On average, completion time was 45 minutes. This is a large demand from any first-time user. Therefore, for an implementation of this explanation strategy in a product or a robo-advisory application, we suggest beginning with a single local or personalised explanation at this stage. An interested user should be allowed options to dive deeper. A sequential reveal of explanations accompanied with

relevant features could be adopted. Additionally, feature effect and behaviour types explanations are highly technical and should be replaced or simplified. Each explanation should clearly state its relevance and intent to each user. Otherwise, user attention is quickly lost.

Effects of explanations on users' trust in explanations, system recommendations

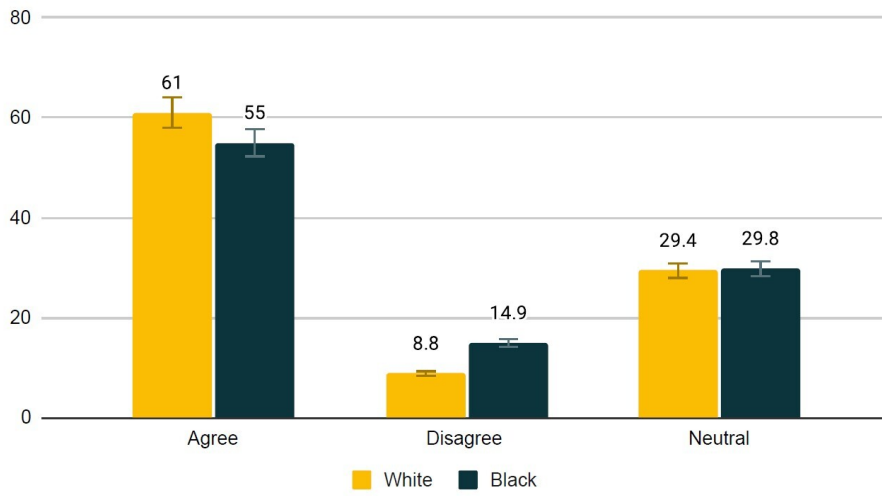
The goal of a subset of qualitative questions from the survey is to understand the effect of comprehension on system user trust. This covers both user trust in the explanations and in the system recommendations. Trust in explanations is determined through measuring trust in the data used for algorithms and the ability of explanations to convey decision making logic behind system recommendations. Trust in the system is measured using direct qualitative questions posed to participants in the survey.

Measuring user trust begins with model specifications which display the sources of data for both algorithms. Data is an essential part of algorithm training and learning. Therefore, a necessary part of user information requirements. A round of qualitative questions is used to gauge participant trust in these external data sources. On average, data sources and variables for fund risk were trusted around 30% more than the ones for user risk determination. Sources for user risk variables were based on other Robo advisory applications in the market which may not be well known. While the sources for mutual fund risk calculations are taken from well reputed investment sites such as AMFI, Mutual Fund India, Money Control etc. Therefore, participants familiar with these sites immediately trust the data. (Observation 13)

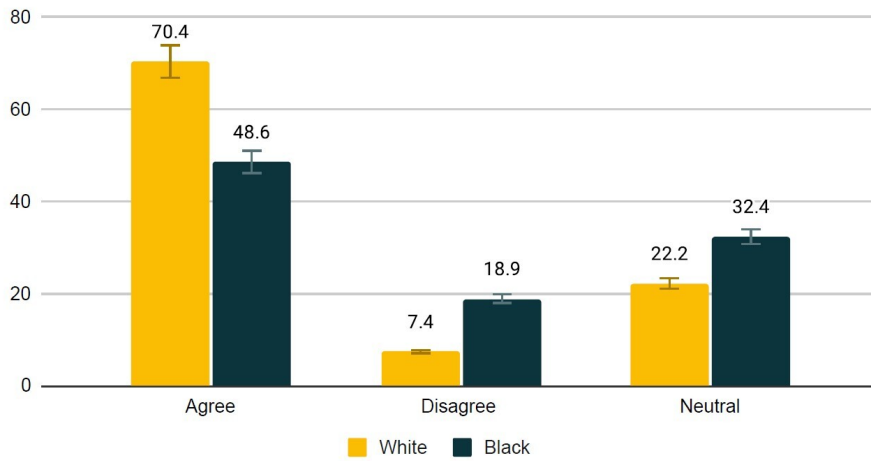
Second, participants were asked directly whether knowing the extent of variables used in the system has an effect on their trust in the system. 70% of white box and 49% of black box participants agreed with this sentiment. Very few white box participants disagreed compared with 19% of the black box disagreements. (Observation 13) This difference occurs due to reduced clarity of explanations due to surrogate approximation and consequently, the reduced user comprehension (refer to observation 5 and 6).

Finally, trust in the *system* was measured by a direct question asking whether users trust the system because of the explanations. 61% of white box and 55% of the black box users agreed. Around 30% of users in each system were neutral – neither agreeing or disagreeing with this statement. Only a small number of users disagreed. This in addition to observation 9, 10 (participants reporting a need for explanations in all decision-making systems) and 12 (an increased general understanding of the system due to explanations) show that presence of explanations would have a positive effect on user trust.

A) User responses to "I trust this system due to explanations provided." (percentage)



B) User responses to "I find that knowing the extent of variables used makes the model more trustworthy." (percentage)



C) User responses to "I find the data and its sources to be trustworthy." (percentage)

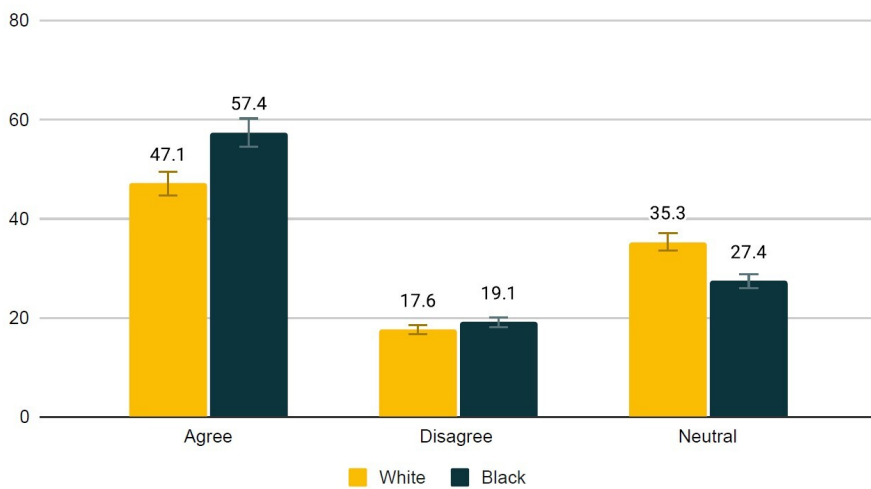


Figure 3. 26 Comparing white and black box explanations for user trust

Figure 3.26 (A to C) shows user responses that help in measuring user trust, comparing between white and black box explanations. It highlights user trust in data sources, variables used in the models and trust in the system due to explanations, represented using three different statements. The graph intends to measure user trust through these various aspects of explanations, based on self-reported scores. The data is normalised to percentages.

We find that the difference between white box and black box explanations is statistically significant (at 95% confidence level)¹⁹ for only one of the three statements — Statement B, “I find that knowing the extent of variables used makes the model more trustworthy.” Here, more individuals agree to the statement in white box explanations than in black box explanations, indicating that users trust the model more often when they know the extent of the variables used.

Difference in effect of explanations based on demographic backgrounds

Participants of the survey come from different backgrounds. This includes a wide range of age groups, risk categories, backgrounds, prior robo advisory or investment knowledge, etc. Three major demographic divisions for robo-advisory users are their risk categories, familiarity with robo-advisory applications, and general characteristics such as age, gender etc. This section analyses the effect of these different backgrounds on usability and comprehension.

Sixty-five percent of users reported either being previously familiar with RA or having used similar services previously. Most of these users (93%) were well acquainted with the variables used for both algorithms. All of these users reported finding data sources for mutual funds trustworthy. While there was no notable difference in comprehension, self-reported understanding of explanations was the highest for this group (7-10/10). Therefore, there is no advantage of background knowledge when it comes to comprehension of

¹⁹ T-test results for all three charts indicating difference in white and black box explanations are available in the Appendix.

explanations. Users can comprehend explanations regardless of being seasoned or amateur.

The next point of difference is between users categorised as high vs low risk. A majority of high risk users belong to the 18-35 age group with stable moderate to high incomes. In contrast, low risk users are usually older, with unstable or fixed lower range incomes. Their investment objectives are usually tax saving or retirement planning. There are no notable patterns in user comprehension or usability. However, users classified into the high risk category prefer being in a comparatively lower risk category. Only two of these users use this discomfort to actively change their answers towards their preference through the robo-advisory questionnaire.

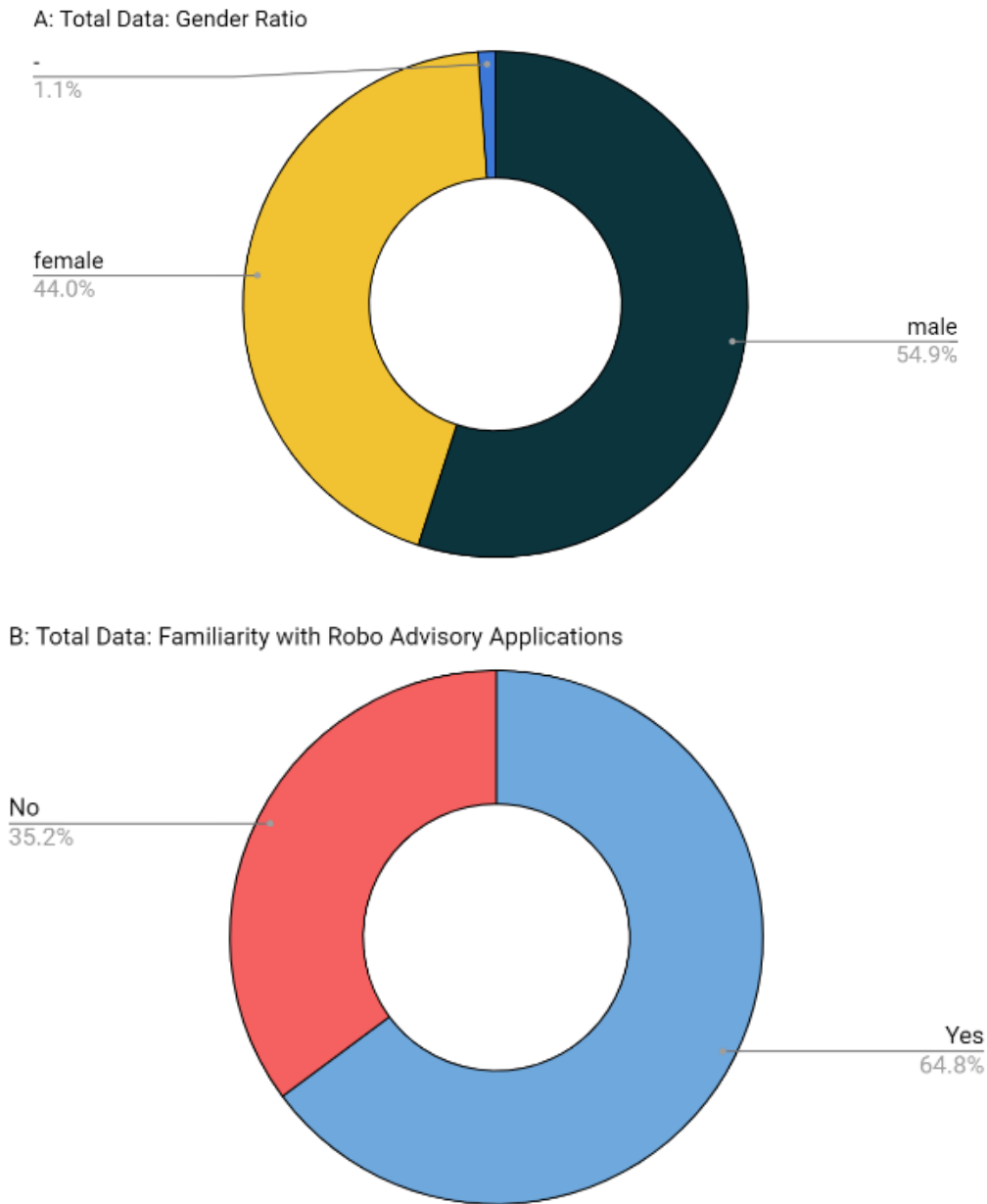


Figure 3. 27 Demographic divisions in users

Figure 3.27 shows major demographic divisions considered in this section based on self-reported genders and familiarity with robo advisors.

Effect of complexity on user comprehension and usability (black box vs. white box)

On average, white and black box comprehension differ by a few percentage points across all explanations (refer to Figure 21), with the exception of feature effects and interactions. As mentioned in observations 5 and 6, the reduction in comprehension is due to the reduced clarity and precision of the explanations due

to surrogate models. However, based on the close comprehension of other explanations, the general ability of participants to interpret technical explanations of the system is comparable.

Consequently, feature importance, effects and interactions were the least popular explanations among black box and most popular amongst the white box participants. Black box participants prefer model specifications, the least favourite of the white box participants. Understanding feature importance is not as important to black box participants (refer to Figure 29). Comparatively, they prefer information of data and sources.

Observation 12 shows an average increase in comprehension, 2 points for white box and 1 for black. Since the comprehensions of both systems is similar but there is a difference in popularity of explanations, a higher preference towards model specifications, we conclude that users can discern the differences between white and black box explanations. Reduced clarity of explanations directly and significantly affects user confidence in explanations. Ideally, users should be provided white box explanations of interpretable algorithms. However, in cases where black box explanations are necessary, the approximations of surrogate models should be explained. Additionally, true explanations should be provided, this could mean revealing and explaining data, features, or sources used by the model to learn.

Broader implications of findings for robo-advisory applications

Based on the participants of the user study an average persona of an interested Robo advisory user emerges. He or she is young, on the lower end of 18 to 55 years. They are currently investing in mutual funds or have their funds saved or in fixed deposits, but they are interested in investing further for wealth creation or monthly income objectives. This average user has 0 to 1 dependent and moderate stable income. While the algorithm marks this user as liking risk, their self-reported answers show moderate risk-taking behaviour and comfort. Therefore, ideally, they should be recommended funds accordingly. This average user, is affected positively by explanations and shows an increased trust in the system.

The user prefers explanations he or she can relate to and understand, such as the feature importance. Although, if interested, the user can comprehend more detailed explanations very well.

A basic explanation relating the user input to the outcome would be effective and most popular (such as local or global feature importance). However, an interested user should be allowed options to understand details such as model specifications and feature effects and interactions. Detailed explanations are highly technical and should be replaced or simplified. Each explanation should clearly state its objective and intent. White box explanations are preferred by robo-advisory users. In case of highly complex algorithms and black box explanations, an external reference through data sources or variables should be provided.

Towards a standardised explanation strategy

Users prefer visual explanations over textual or tabular information. Our first observation showed around 30% of users responding better to visual explanations, rather than tabular or textual, based on findings (Observation 1, 4, 13). Therefore, the most essential aspects of the system should be conveyed through visualisations made as interactive as possible. However, global explanations are as effective as local, that is, highly personalised explanations are not required.

Systems that require user input should communicate the extent and need of data requirements to users. the release of this information should be carefully done as it could have negative consequences. Such information should be released after understanding any legal and ethical constraints. An ambitiously fair and effective plan for data collection from users could involve providing users with the ability to consent incrementally. For example, in our case, this could involve gathering only a few parameters of the robo-advisory questionnaire, whichever ones the user feels comfortable revealing; such as age or investment objective. The user could be made aware of the limitations caused by the partial status of their form. Where they could be shown funds that similar users of that age and objective invest into. This would allow users to opt out of revealing certain data at the expense of performance.

Every AI prediction is based on data, so data sources need to be part of explanations. Explaining data and sources is important to provide users with additional context. Observation 13 and Figure 29 show that users trust is affected positively after revealing data sources. Since data are the fundamental requirements of any AI-based ADS, the quality of data influences the outcome. Training algorithms on data with missing samples or biased groups can lead to problematic predictions when deployed for practical use. COMPAS (Julia Angwin, 2016) recidivism predicting algorithm had biased data for minority groups resulting in a biased prediction. Image recognition software trained on gender-biased images is often 'sexist' (Simonite, 2017). Thus, transparently explaining at least a few details regarding this impactful data is necessary, especially in a high stakes fintech recommendation algorithm.

Keeping in mind data as an asset, our suggestion is to disclose these details only to an official regulator. After all, such descriptions and warnings of data are common across sectors such as pharmaceuticals (for FDA approval (James Woodcock, Center for Drug Evaluation and Research, 2020) machinery (for safety/ security reasons), etc. The system depicted in the paper is simple compared to most products in the market that collect user data. In such cases, users can be surprised by the extent of their data being used. Therefore, the user onboarding and explanation revelations should be done gently and in stages.

User expectations of the outcome, recommendations, or predictions should be explained clearly and directly. AI is capable of providing invaluable insight for many domains; from designing drugs and pharmaceuticals to diagnosing patients with FDA approval. However, overpromising AI results is a problem that can affect users on a personal level. An instance where promises have fallen short are IBM Watson, which set out to revolutionise the health care industry after its win on the game show Jeopardy in 2014. While the win was significant technologically it was heavily criticised for over promising and under delivering. Similarly, many AI products make promises about the product without delving into exactly what the AI service predicts. Over or under promising the effectiveness of predictions or recommendations is confusing for users. This also increases liability and risk;

the explanations provide a way to manage user expectations by revealing the probabilistic working logic. Therefore, the scope and reach of the AI system should be effectively communicated. User expectations from both the AI system as well as the explanations need to be managed.

A way to increase user confidence in prediction is by conveying confidence scores to each user. Each model has a confidence value associated with each classification or recommendation. Our system explained this through localized explanations. While local explanations showed high level of comprehension from users, the effect of showing confidence values varies per user/domain/complexity. It is difficult to evaluate its direct effect on every user. Most prediction algorithms have a threshold beyond which a prediction is positive and under which it is negative (for our algorithm this could mean the difference between classifying a user as low risk or moderate risk). These confidence intervals are interpreted as probabilistic values (most common threshold is 50%), users could misinterpret lower confidence predictions (such as 70% or 80%) to imply uncertainty or a limitation of the algorithm. Therefore, such explanations should be communicated with care or only in high risk or high reward areas.

While explaining internal logic to users of our system, many initially struggle to form an accurate mental model of an AI-powered product because the way these systems execute tasks is different from the way a person would. Therefore, users often confuse the decision-making logic of AI with the way they would personally solve problems. To an extent, this disconnect can be bridged through explanations, but may cause the users to develop a negative or positive mental bias. For instance; a user may be able to understand the sequential logic of an algorithm but it is still impossible to comb through millions of datapoints to grasp the countless patterns observed by a machine. For example, neural networks, today, commonly make use of over a million nodes and functions to transform data. The key is to communicate the system's limits and capabilities in a way that doesn't create or support extreme expectations. This could be product or domain dependent. Additionally, explanations can be used to explain the benefits of the system rather than focus on the technology intricacies.

Conclusion and Future Work

In conclusion, we designed and implemented a framework for user-centric and multi-perspective explanations of a robo financial advisory application based on an interdisciplinary review of existing tools, methods, and research. We identified the goals and information needs as expected by users from explanations of complex decision support systems. We examined the usability of the multi-perspective framework of explanations, and the broad effects this has on user trust and system usability through a user study with human subjects. Our experiment demonstrates the usefulness of such types of explanations from the perspectives of both novice and seasoned investors. Additionally, we differentiate between white and black box explanations. We find that black box explanations of the same system provide an imprecise explanation that leads to reduced user comprehension and confidence in the system. Our major finding suggests that users are well equipped to understand explanations of a complex algorithm, particularly explanations that provide a personalised but partial overview of how the system uses features. Additionally, these explanations are found to be positively correlated with user trust and consequently usage of the system. Therefore, explanations of decision logic would benefit users as well as the developers of these systems. We are also able to provide an average persona of the users interested in robo-advisory applications, along with their objectives, hopes, behaviour and comforts.

There are a number of avenues of future work that we would like to explore. Our next project works on measuring the difference of usability and trust between white and black box explanations. Additionally, we wish to use our insights to design a more generalisable strategy of explanations, complemented by cross-domain (legal, ethical, and policy) support.

Appendix

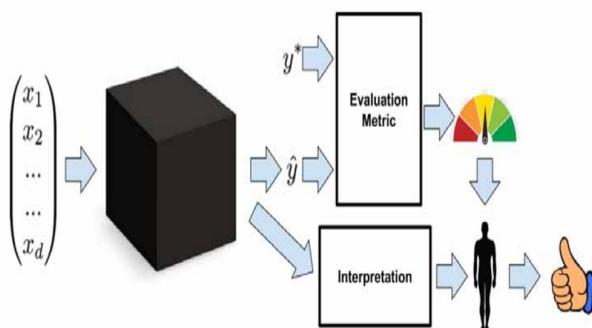
What is XAI? What is the need for it?

AI is increasingly omnipresent, from Google's Smart Compose (Chen, 2019) designed to assist in writing mails to self-driving cars (Mayank Bansal, 2018) on the streets. With the fast and widespread adoption of artificial intelligence (AI), we are shifting towards a more algorithmic society. AI's rising popularity is due to its remarkable ability to perform complex tasks. These algorithms learn intricate details and patterns, that humans cannot, from large quantities of high-dimensional data. As AI models get increasingly complex, growing from using simple linear logic to complex multilayer hybrid networks there is reduction in visibility and knowledge on how AI systems make the decisions. As a result, the knowledge gap between users, decision-makers and system architects keeps getting wider. (Shane Mueller, 2019)

Many of the algorithms used for machine learning are deemed as "black boxes", which means that they cannot be examined in a bid to understand the "how" and the "why". Indeed, the black box nature of these systems allows for powerful predictions, but lacks transparency and accountability. The effectiveness of a transparent model rests on simulatability (possible to understand & easily predict output), decomposability (possible to unpack more at the level of single components) and algorithm transparency (possible to decipher level of training algorithm). This is one of the main barriers preventing many practical applications of AI. This issue has triggered the need for explainable AI (XAI) (Shane Mueller, 2019). XAI holds significant promise for improving the trust and transparency of AI-based systems. Explainable AI is artificial intelligence programmed to describe the decision-making process in a way that is understandable to humans (Alejandro Barredo Arrieta, 2019). As described by D. Gunning (Shane Mueller, 2019) "XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners".

All algorithms are tested for certain performance metrics; such as F1 Score, AUC etc to determine their usability. Unfortunately, using simple test performance

metrics isn't enough to prove whether the algorithm will hold up in real life. Even (Lipton, 2018) argues that the demand for a human understandable explanation arises when there is a mismatch between the formal goals of supervised learning (test set predictive performance) and the real-world costs in a deployment setting. XAI has the potential to become a reliable way to gauge how the model makes decisions i.e., how it would make decisions when deployed practically. Thus 'explainability' should be a requirement along with performance metrics, especially for user centric ADS.



Appendix Figure 3.1. Evaluation metrics require only predictions (\hat{y}) and ground truth labels (y^*). When stakeholders additionally demand interpretability, we might infer the need for additional metrics that are not captured

currently. Source: (Lipton, 2018)

To put it in clearer terms, the goal of XAI (Hoffman, 2018) is to inspect an algorithmic system in order to understand the steps and models involved in making decisions, by asking and addressing some specific questions such as: why an AI system makes a specific prediction or decision? Why doesn't the AI system do something else? How to identify when the system is affected by a change in the input? When does the AI system succeed and when does it fail? When do AI systems give enough confidence in the decision that you can trust it, and how can the AI system correct errors that arise?

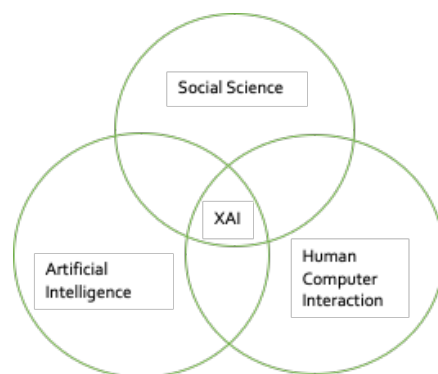
Artificial intelligence has become a driver of increase in productivity as well as societal change in the mainstream governmental discourse across several countries (Amber Sinha, 2018). The demand for XAI has increased further, with calls for algorithmic regulation from the social sectors. Answering these questions is critical to assuage certain ethical, policy and legal anxieties and

concerns that surround ADS. Illustrated in the table below are various studies that bring to light the need for explanations and the complexity of these concerns surrounding the use of ADS. The concerns surround the notorious 'black box' nature of AI algorithms; meaning their decision making is even unknown to the developers. This simply cannot do as recent examples of the Apple cards' credit limit discrimination between men and women (BBC, 2019) , minority bias of the recidivism predicting COMPAS (Julia Angwin, 2016) algorithm and the scrapped tools designed by Facebook and Google that preferred male software engineers - prove that AI can be risky, especially when it affects sensitive groups of society.

Author	Title(s)	Concerns surrounding algorithmic systems
(Holdren, 2016)	Preparing for the Future of Artificial Intelligence	The report delves into a range of issues concerning AI application, calling for accountability, fairness and transparency through some form of explanation.
(Bathee, 2018)	The AI Black Box and the Failure of Intent and Causation	Argues that current legal systems are ill-equipped to deal with the legal issues surrounding ADS accountability and fairness. The report argues transparency is an AI problem, and suggests solving it by constructive detailed explanations of the system.
(Stanford, 2009)	One Hundred Year Study on Artificial Intelligence	Acknowledges that AI may cause problems with civil and criminal liability doctrines, such as intent, privacy, labour and innovation policy. Promotes increased system interpretation and explanation via human interactions to build trust and prevent drastic failure.
(Sandra Wachter, 2016), (Bryce Goodman, 2016)	European Union General Data Protection Regulation	Algorithmic regulation should aim to increase transparency and accountability. Firmly states that humans have a right to non-discrimination, right to explanation and a right to information, when it comes to algorithms as well.
(CIFAR, 2017), (Singapore, 2018), (MEITY, 2019)	AI Strategy	Concerns raised on AI's effect and impact on safety, ethics, privacy, fairness, trust, transparency, accountability, interpretability, bias, fairness and usability. Demand accountability and transparency through explanation, similar to the policy defined by EU.

Appendix Table 1: Global demand highlights the need for XAI and urgency for increased transparency, accountability and interpretability of complex algorithmic decision-making systems.

To sum up, due to the cross-sectoral and cross-agency (regulatory/legal, policy, development etc.) concerns regarding ethics, security, safety, accountability and interpretability of ADS, XAI is beneficial to a diverse set of people involved in the ADS development and deployment process (Alun Preece, 2018). These people include developers of AI; computer scientists, social scientists, academicians, businessmen – all the people involved in building AI applications. For them, explainability is a way to gain assurance of the good quality of a system. The clarity provided by XAI is also a relief for ethicists; an interdisciplinary community of people, mainly concerned with fairness, accountability, and transparency of AI systems, including but not limited to policy-makers, journalists, economists, politicians, commentators, and critics. Lastly, it is absolutely beneficial to the users of AI systems, who require instructions on the purpose, appropriate usage and expected results in order to understand and trust these systems. Inexplicable algorithms face higher chances of rejection during practical deployment, in spite of being highly accurate, hindering innovation and slowing down scientific progress. Therefore, **XAI is crucial at this stage of the AI evolution.**



Appendix Figure 3.2: Interdisciplinary scope of XAI

The desired properties of the XAI systems include informativeness, low cognitive load, usability, fidelity, robustness, non-misleading and conversational/interactive.

Literature Review on Measuring Effectiveness of Explanations

Numerous studies have found that intelligent systems are trusted more if their recommendations are explained and numerous studies have linked trust to explanation in the context of AI systems use. (Hoffman, 2018) provides a review on this topic.

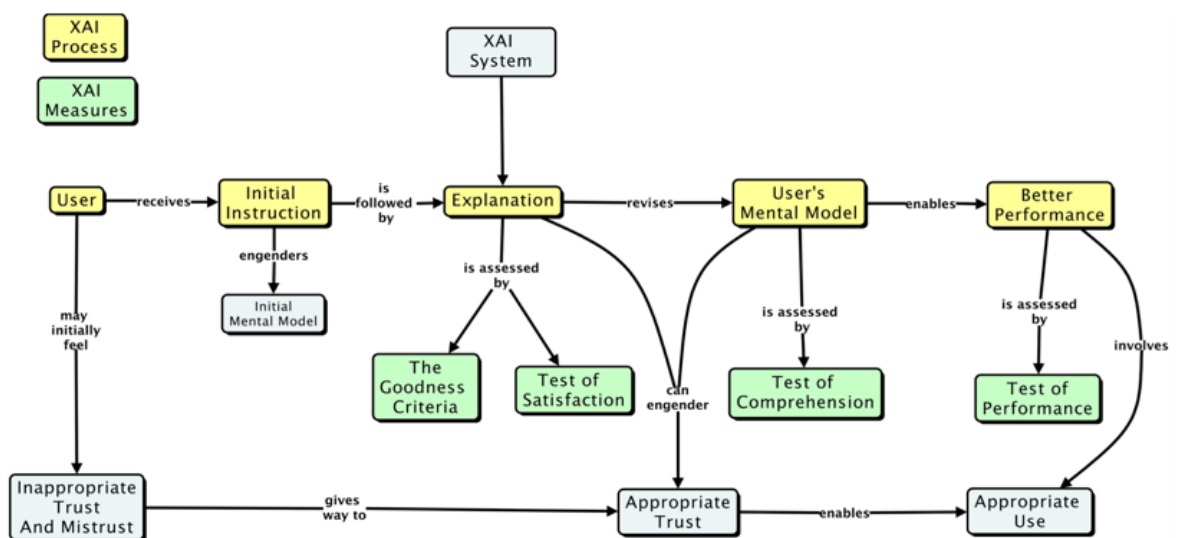
A tremendous body of research has covered the necessary aspects of an explanation. (Or Biran, 2017), (Shane Mueller, 2019). (Miller, 2018) conducted a vast survey of publications on explanations. An interesting point made by Miller's paper is that XAI research does not build upon existing techniques used in HCI or UX research. He argues that "The very experts who understand decision-making models the best is not in the right position to judge the usefulness of explanations to lay users". In another "graspability" test (Kim, 2018) identifies the need to differentiate between 'good' and 'correct' explanations. A correct explanation will be proven with performance metrics while a good explanation covers correlation and/or causation, counterfactual reasoning, assumptions and factual revelations. Thus 'good' explanations should be designed using the fields of philosophy, psychology, cognitive science merged with XAI.

(Miller, 2018) states that explanations are usually short answers to 'why' questions. Additionally, good explanations are contrastive or relative. Explanations rarely consist of an actual and complete cause of an event. Explanations are meant to be a social transfer knowledge, presented as part of a conversation or interaction, and are thus presented relative to the explainer's beliefs about the explainee's beliefs. Causal relationships convey information in a human understandable manner as opposed to referring to probabilities or statistical relationships in explanation.

Covering all the concepts about a user experiment, to evaluate the effectiveness of explanations, is conducted in this study. The use case for this study is a robo advisory mutual fund recommendation algorithm. A replica system was designed for the users to experiment, based on suggestions from XAI-HCI research. The primary design concept adopted is the creation of mental model and its

evaluation (Hoffman, 2018). Mental models are a popular way to **enhance user understanding of complex systems**. (Hoffman, 2018) identifies a few essential attributes of explanations: understandability, feeling of satisfaction, sufficiency of detail, completeness, usefulness, accuracy, and trustworthiness.

Based on the conceptual model of the XAI explaining-evaluation process presented below, the replica robo advisory system uses four major classes of measures. First, initial instructions on how to use an AI system are given. These enable the user to form an abstract mental model of the task and the system. Subsequent experience, which includes the white and black box system-generated explanations, enables participants to refine their mental model, which should lead to better performance and appropriate trust and reliance. These guided tasks can elicit desired mental models in system users' minds (Nancy Stagers, 1993).



Appendix Figure 3.3: A conceptual mental model approach to the explanation process. (Hoffman, 2018)

This approach is accompanied by many questions asked by the users internally, after running them through the entire process of the system (Miller, 2018). By asking these questions in the user study, the explanation quality and usability are determined by whether it fulfils four essential requirements of the user. Questions and user requirements answered are listed below:

5. **Information requirements:** required knowledge to provide an adequate explanation.
6. **Information access:** justifications, what information the explainer has to give the explanation such as the causes, the desires, etc.
7. **Pragmatic goals:** refers to the goal of the explanation, such as transferring knowledge to the explainee, making an actor look irrational, or generating trust with the explainee.
8. **Functional capacities:** each explanatory tool has functional capacities that constrain or dictate what goals can be achieved with that tool.

Once the users are satisfied with an understanding of the system procedure and goals, they turn towards the explanation of the algorithms used in the process. To evaluate how effectively the explanation communicates its intended purpose, we use a questionnaire-based approach created by (Andreas Holzinger, 2020) termed as the System Causability Scale. The user study contains a few questions inspired by SCS.

To provide relative/contrastive explanations, we draw from the Quantitative Input Influence strategy (Anupam Datta, 2016). Relative comparison between similar groups in data; across different cultural, demographic, or phenotypic groups (e.g., race, geographic location, sex) and intersectional groups (e.g., age and race, or sex and age). This allows an additional insight into abductive, counterfactual and contrastive reasoning compared to other absolute measures. Following the mental model approach, users are given explanations of the variations of groups and data used to train the models. After mental modelling questions, they are shown answers for individual outcomes (why did a particular user get this?), group outcomes (What did users similar to you get?), group disparity (What is the difference between two groups?). The user study, along with its explanations and questions, is further elaborated in the results section.

Robo Advisor: User Risk Questionnaire

1. What is your age (in years)?
2. How many people depend on you financially?

3. What is your annual income range?
4. What % of your monthly income do you pay in outstanding loans, EMI etc?
5. Please select the monthly stability of your income.
6. Where is most of your current portfolio parked?
7. What is your primary investment objective?
8. How long do you plan to stay invested?
9. To achieve high returns, are you comfortable with high-risk investments?
10. If you lose 20% of your invested value one month after investment, you will?

Robo financial advisors: going into the details

Robo advisory (RA) applications are automated web-based data-driven investment advisory algorithms that estimate the best plans for trading, investment, portfolio rebalancing, or tax saving, for each individual as per their requirements and preferences. RA's have evolved from simple questionnaire-based suggestions to fully automated systems that cover entire investment/portfolio management process using quantitative methods and algorithms. Core algorithms perform asset allocation and portfolio optimisation (Deloitte, 2016 (a)) (Deloitte, 2016 (b)).

Typically, a user fills in questionnaire or survey and is classified in either three or five risk classes (ranging from 'low risk' to 'high risk'). RA's open up the potential for finance to be democratised by reducing the financial barrier to entry and providing equal access to financial advice through their low-cost business model (Laboure & Braunstein, 2017). Moreover, unlike human advisors, RAs provide no reasons or explanations for their decisions, and this shortcoming reduces the trust that users repose in their advice (Maurell, 2019). The RA output is often a set of recommendations for allocations based on parameters like the size of funds (small, mid-cap), the type of investment (debt and equity funds), and even a list of securities or portfolios (IOSCO, 2014). The predicted reach of the RA market by 2022 is 2.2 trillion (Matt Thomas, Andy Masters, 2016).

After creating a representative set of risk efficient instruments covering different asset classes and types, risk profiling is done by robo financial advisors. This is a

crucial step to determine the risk class of the user which in turn determines the investment advice. Irregularities at this primary step will lead to incorrect recommendations for the users. This profiling included investment goals, age of investor, dependents etc. For details refer to (Krishnan, et al., 2020) Once a risk profile is assigned asset allocation is determined using advanced quantitative methods.

The main reasons for the success of RA's are (Insider, 2017):

1. A new generation of educated, digitally savvy clientele.
2. Low cost of RA services compared to traditional financial advisors, minimal starting investment, the ability to deal with large amounts of data and multi-device flexibility.
3. Concentration of global wealth and adoption of RA's in Asia.

A few examples of RA firms are; Betterment, WealthFront, SigFig, Ellevest, Ally, Charles Schwab. Several robo financial advisory applications operate in India alone. Prominent ones include PayTM money, GoalWise, Artha-Yantra, Upwardly, Kuvera, Scripbox, MobiKwick, RoboAdviso, 5Paisa.com, ETMoney, FundsIndia and Tavaga, among others.

Evolution of explainable AI (XAI)

First Generation Systems

(1970 - 1985)

Explainable AI started with **Expert Systems** for decision aiding and diagnosing. These systems lacked user-friendliness. Therefore, faced pushback from domain experts; such as medical practitioners. The issue was that these systems gave final predictions without providing a rationale (William R. Swartout, 1988). (Mooney, 2005) found that explanations of how computer systems work could convince users to adopt the system recommendations, but it does not imply that users will be satisfied with their decisions.

The *global explanation versus local justification* theory conducted research on how expert systems could explain their decisions in a user understandable manner (Chandrasekaran, 1989).

Explanations help **users trust** the system but the effect is **dependent on the skill of the users** (Donna Lamberti, 1990). Leading to a rise in **Explanation Systems** in the 1990's who used different methodologies to form logical or probabilistic equations to explain decisions made by the system. Decisions in the systems were made using "knowledge" or rule sets provided as an explanation to a layperson. It was found that these explanations provided more clarity to developers and were quickly adopted for this purpose (Michael R. Wick, 1992). First-generation systems **could not justify inferences**; why in a certain situation a particular interpretation or action was correct. (Clanccy, 1981). Further simply **'recapping' the internal workings of the system did not provide explanations considered useful by users** or domain experts of the system (Swartout, 1993).

During the second generation, researchers that effective explanations describe **why choices were not** made, showing **alternative courses** of action, or infer that a specified behaviour **differs from typical** by establishing baselines. This involves **contrastive reasoning**, which has been shown to be essential for human learning (Lombrozo, 2010). Another technique researched was **concept mapping**, creating 'knowledge and inference' directed graphs that represent the reasoning of the algorithm. These interactive maps were shown to users as explanations.

Third generation (2015+)

Research into XAI is on the rise, possibly due to an increase in big data systems and ubiquitous computing. In this era of explanation, a large number of research papers focus on **user understanding** by visualising, comparing, and understanding the operation of black box algorithms. A number of researchers have focused more specifically on explanations that might be presented to non-developer users of the system. (James Wexler, 2019), (Shane Mueller, 2019), (Avanti Shrikumar, 2017), (Marco Tulio Ribeiro, 2016), (Sai P. Selvaraj, 2018)

Increased importance is placed on defining and **standardising XAI** taxonomy and concepts such as; understandability (G. Montavon, 2017), interpretability, transparency, fairness, comprehensibility, accountability (Ananny, 2015). Possibly a result of increase in regulatory and legal attention into data governance. Another interesting trend is the shifting purpose of explanations; **increasing trustworthiness** (Lipton, 2018), (Došilović, 2018), (Doran et al., 2018) of domain experts and users of these systems. Lately, XAI and HCI research intensified focus on focuses on user interaction, usability, practical interpretability, and system efficacy. (Jichen Zhu, 2018), (Patrick Hall, 2019)

Table on sampling strategy

Sequence	Sampling Strategy (primary feature)	Further divisions
Iteration 1	25-35 age range	High income (20%), Middle income (40%), Low income (40%)
Iteration 2	35-55 age range	High income (40%), Middle income (40%), Low income (20%)
Iteration 3	55+ age range	High income (40%), Middle income (40%), Low income (20%)
Iteration 4	Novice/ Beginners	Mixed age groups.
Iteration 5	Seasoned/ Experts	Mixed age groups.

The table above describes sampling strategy used to choose users along with divisive features chosen to select groups. Two different groups of similar iterations are used for black box and white box explanations.

User Survey Contents

The table describes explanation types along with logic conveyed by these explanations and the specific methodologies used to present these explanations. To understand the extent of user understanding questions have been designed to specifically cover each aspect.

Explanation Type	Intentions	Relevant qualitative and quantitative questions
Robo Advisory Questionnaire	Gather participant backgrounds and preferences for user risk calculations and fund advice.	<ol style="list-style-type: none"> 1. What is your age (in years)? 2. How many people depend on you financially? 3. What is your annual income range? 4. What % of your monthly income do you pay in outstanding loans, EMI, etc? 5. Please select the stability of your income 6. Where are most of your current portfolio parked? 7. What is your primary investment objective? 8. How long do you plan to stay invested? 9. To achieve high returns, you are comfortable with high-risk investments 10. If you lose 20% of your invested value one month after investment, you will? 11. To which gender do you identify most? 12. Are you familiar with robo advisors and/or have used them before?
Localized Explanations	Provide a view of the effect of personalized preferences on recommendations.	<ol style="list-style-type: none"> 1. Which feature has the most impact on my risk decision? 2. What happens to my risk category if I change my Age? 3. On a scale of 0 to 10, how confidently have you understood the models' internal decision-making logic? 4. On a scale of 0 to 10, how much do you agree with the risk category assigned to you?

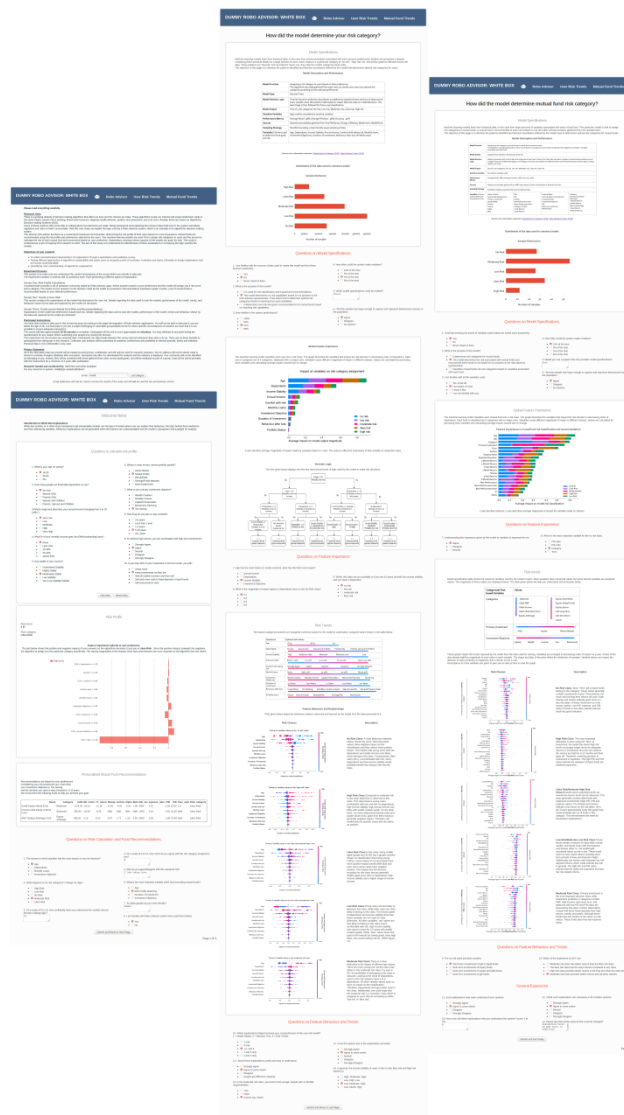
		5. Why do you agree/disagree with this assigned risk?
Mutual fund recommendations.	Provides a personalized explanation for fund recommendation.	1. What is the most important variable while recommending mutual funds?
Model specifications for user risk.	Explains information aggregation and requirements.	1. I am familiar with the sources of data used to create this model and find these sources trustworthy. 2. What is the purpose of this model? 3. How reliable is the system performance? 4. How often could the system make mistakes? 5. Which model specifications could be omitted? 6. I find the sample size large enough to capture and represent behavioural nuances in the population.
Global feature importance for user risk	A holistic view of the order and magnitude of feature importance	1. Age has the most impact on the model outcome, what has the third most impact? 2. What is the magnitude of impact (approx.) dependents have on the No-Risk class?
Decision Flow	A hierarchical view of logic	1. Which risk class are you probably in if you are 44 years old with low-income stability, and you have 1 dependent
User risk model feature effects and behaviour	Feature interaction, behaviour, positive and negative impacts.	1. In the moderate risk class, you tend to find younger people with no familial responsibilities. 2. In general, the income stability of users in the no risk, likes risk and high-risk classes is: etc.
Model specifications for fund risk.	Explains information aggregation and requirements for	What is the purpose of this model? 17. How often could the system make mistakes?

	more complex algorithms.	19. I find the sample size large enough to capture and represent behavioural nuances in the population.
Global feature importance for fund risk	A holistic view of the order and magnitude of feature importance	1. What is the most important variable for the no risk class?
Fund risk model feature effects and behaviour	Feature interaction, behaviour, positive and negative impacts.	1. Which of the statements is NOT true. 2. What are the primary contents of the no risk class? 3. The moderate risk class depicts the following behaviours...
User Opinions		1. I find that knowing the extent of variables used makes the model more trustworthy. 2. Understanding the importance given by the model to variables is important for me 3. I think such explanations are necessary in all complex systems. 4. I found these explanations useful and easy to understand. 5. I trust this system due to the explanations provided.
Usability of Explanations	Qualitative questions to determine the usability of explanations.	1. Which explanations helped increase your comprehension of the user risk model? 1 = Model Details, 2 = Decision Tree, 3 = Risk Trends
Usability of System		1. Would you use a system that only provides model specifications? 2. How much did these explanations help you understand the system? (score 1 to 10)

Appendix Table 2: User survey questions.

Complete overview of the white box system

Figure below show the samples of the screens used for white box user study.



Appendix Figure 3.4: White Box User Study Screens

Mutual Fund Quantifiers: fund performance, risk, and recommendations

Choosing parameters for the mutual fund algorithm.

Fund Specific Parameters

Explanations of these parameters give investors insight into the specific funds' performance.

Alpha A measure of an investment's past performance on a risk-adjusted basis. Alpha is used to determine the performance of a portfolio, compared to a benchmark index. Alpha of 1

means that the fund has outperformed the benchmark by 1%. Ideally a fund should have a high alpha value.

Beta	Measures the volatility of the fund compared to the market. Beta of 1.20 implies the stock is 20% more volatiles than the market.
Sharpe	Determines whether an investment's returns are due to wise investment decisions or the results of excess risk.
Net asset value	Represents a fund's per-share market value. NAV is calculated by dividing the total value of all the cash and securities in a fund's portfolio, minus any liabilities, by the number of outstanding shares.
Sortino	Differentiates harmful volatility from total overall volatility by using the asset's standard deviation of negative portfolio returns, called downside deviation, instead of the total standard deviation of portfolio returns.
Treynor	Measure of the returns earned more than the risk-free return at a given level of market risk. It highlights the risk-adjusted profits generated by a mutual fund scheme.

Peer performance parameters: Explanations of these parameters give comparative insights into the funds' performance.

Mutual fund holdings	Represent securities (stocks or bonds) held in the fund. Apart from conveying necessary details of the investment, number of holdings also determine the funds risk. Fewer holdings could mean volatility and risk can be significantly high because there are fewer holdings with a larger impact on the performance of the mutual fund. Conversely, if a fund has is large then its performance is likely to be like an index.
Price to earnings ratio (P/E)	Interpreted as the amount of time required to get a return of investment. It is calculated by dividing the market price of a share by the earnings per share.
Price to book ratio (P/B)	Used to compare firms market capitalisation to its book value.
Asset under management	Overall market value of assets/capital that a mutual fund holds. AUM conveys the funds popularity. Increased popularity implies increase in investment and market value. AUM is heavily affected by market fluctuations but it is an informative comparative parameter.
Best Performance year	Best performance shown by stock in all years of operation.
Worst Performance year	The worst performance shown by stock in all years of operation.

Benchmarks are index funds against which fund performance is measured.

Expense Ratio: Robo advisors usually have very low expense ratios, these are management fees charged by fund manager. Robo advisors such as Betterment, Vanguard, Wealthfront etc. charge fees in the range of 0.4 to 0.0 % of deposit, annually (Ludwig, 2020).

Appendix Table 3: Elaboration of few dimensions/ Variables used in mutual fund risk calculation.

Mutual Fund Types: recommendations

The types of mutual funds recommended through the system. Detailed purpose of the funds and primary investments.

Mutual Fund Type	Definition
Equity Funds	Invest in stocks and are the riskiest kind of mutual funds. Investment objectives could be capital accumulation or wealth generation. Large cap mutual funds are ideal for beginner, low-moderate risk investors. While small-mid cap equity funds provide high returns for risk savvy investors.
Debt Funds	invest in fixed income short- or long-term bonds. Focused on capital preservation and growth, these are low capital and low risk funds.
Balanced Funds	contain a mixture of debt and equity funds. Their goals lie between income and capital appreciation. Ideal for conservative, retired investors looking for long term investments.
Liquid Funds	Short term and high liquidity investments. Ideal for high capital, no risk investors looking for income and capital preservation for a few days to months. They contain bank fixed deposits, treasury bills, commercial papers, and other debt securities with maturities up to 90 days.
Gilt Funds	invest primarily in government securities with medium to long term maturity. The goal of these funds is wealth accumulation. It is ideal for low risk investors.
Dynamic Funds	are debt mutual funds containing short- and long-term debt bonds. The goal is capital preservation and growth in rising and falling market scenarios. Ideal for moderately risk investors who can wait 3-5 years.
Funds of Funds	Invest in other schemes of mutual funds. Ideal for smaller, moderate capital and low risk investors. The focus is on long term wealth creation.

Appendix Table 4: Fund types

Algorithm Performance Details

For a comparative study on algorithm selection and determination, refer to (Krishnan, et al., 2020).

T-test results

1. T-test results for Figure 3.23

```
. ttest whitebox == blackbox, unpaired unequal
```

Two-sample t test with unequal variances

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
whitebox	9	64.55556	4.799048	14.39714	53.48893	75.62218
blackbox	9	52.11111	5.926505	17.77951	38.44457	65.77766
combined	18	58.33333	3.995095	16.94975	49.90442	66.76225
diff		12.44444	7.625898		-3.778734	28.66762

```
diff = mean(whitebox) - mean(blackbox)      t = 1.6319
Ho: diff = 0                                Satterthwaite's degrees of freedom = 15.3369

Ha: diff < 0                                Ha: diff != 0                                Ha: diff > 0
Pr(T < t) = 0.9385                          Pr(|T| > |t|) = 0.1231                          Pr(T > t) = 0.0615
```

Given the alternate hypothesis that the difference between the mean of white and black box explanations > 0 , the probability $Pr(T > t) = 0.0615$ indicates that we can reject the null hypothesis (diff = 0) with 90% confidence.

2. T-test results for Figure 3.26 (A)

Two-sample t test with unequal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
black	47	.4042553	.1082173	.7419005	.1864252	.6220855
white	34	.5294118	.1135656	.6621953	.2983609	.7604626
combined	81	.4567901	.0786644	.7079792	.3002431	.6133372
diff		-.1251564	.1568698		-.4376239	.187311

```
diff = mean(black) - mean(white)      t = -0.7978
Ho: diff = 0                                Satterthwaite's degrees of freedom = 75.4878

Ha: diff < 0                                Ha: diff != 0                                Ha: diff > 0
Pr(T < t) = 0.2137                          Pr(|T| > |t|) = 0.4275                          Pr(T > t) = 0.7863
```


Here as well (same as Point 2), we failed to reject the null hypothesis that there is statistically significant difference between users of white and black box explanations (i.e., no statistically significant difference observed).

References

- Aaron Fischer, C. R. a. F. D., 2018. Model Class Reliance: Variable importance measures for any machine learning model class, from the 'Rashomon' perspective.. <http://arxiv.org/abs/1801.01489> .
- Abdul, A. V. J. W. D. L. B. Y. & K. M., 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems.. *CHI Conference on Human Factors in Computing Systems - CHI '18*..
- Alejandro Barredo Arrieta, N. D.-R. J. D. S. A. B. S. T. A. B. S. G. S. G.-L. D. M. R. B. R. C. F. H., 2019. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *arXiv:1910.10045 [cs.AI]*.
- Alun Preece, D. H. D. B. R. T. S. C., 2018. Stakeholders in Explainable AI. *Presented at AAAI FSS-18: Artificial Intelligence in Government and Public Sector, Arlington, Virginia, USA*.
- Amber Sinha, E. H. a. A. B., 2018. AI in India: A Policy Agenda. *The centre for internet and society*.
- AMFI, n.d. *Association of Mutual Fund India*. [Online]
Available at: <https://www.amfiindia.com/>
- Ananny, M., 2015. Towards an ethics of algorithms: Convening, Observation, Probability and Timeliness. *Science, Technology and Human Values*.
- Andreas Holzinger, A. C. & H. M., 2020. Measuring the Quality of Explanations: The System Causability Scale (SCS). *Springer*.
- Andreas Holzinger, C. B. C. S. P. D. B. K., 2017. What do we need to build explainable AI systems for the medical domain?.
- Andrej Karpathy, L. F.-F., 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions.
- Angwin, J., Larson, J., Kirchner, L. & Mattu, S., 2016. Machine Bias. *ProPublica*.

Ansgar Koene, C. C. Y. H. H. W. M. P. C. M. J. L. D. R., 2019. A governance framework for algorithmic accountability and transparency. *Panel for the Future of Science and Technology*.

Anupam Datta, S. S. Y. Z., 2016. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems.

Avanti Shrikumar, P. G. A. K., 2017. Learning Important Features Through Propagating Activation Differences. *PMLR 70:3145-3153*.

Bahrammirzaee, A., 2010. A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems. *ACM*.

Bank, W., 2019. *Robo-Advisors: Investing through Machines*. [Online]

Available at:

<http://documents1.worldbank.org/curated/en/275041551196836758/pdf/Robo-Advisors-Investing-through-Machines.pdf>

Bathee, Y., 2018. The Artificial Intelligence Black Box and the Failure of Intent and Causation. *Harvard Journal of Law and Technology*.

BBC, 2019. *Apple's 'sexist' credit card investigated by US regulator*. [Online]

Available at: <https://www.bbc.com/news/business-50365609>

[Accessed 07 04 2020].

Been Kim, C. C. a. J. S., 2013. Inferring robot task plans from human team meetings: A generative modeling approach with logic-based prior. *Association for the advancement for artificial intelligence*.

Been Kim, M. W. J. G. C. C. J. W. F. V. R. S., 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). *ICML*.

Been Kim, M. W. J. G. C. C. J. W. F. V. R. S., 2018. *Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)*. s.l., s.n.

Beketov, M. L. K. & W. M., 2019. Robo Advisors: quantitative methods inside the robots.. *J Asset Manag* 19, 363–370 (2018)..

Bekker, L., 2000. User Involvement in the Design of Human—Computer Interactions: Some Similarities and Differences between Design Approaches.

Best, M. a. R. G., 1991. On the sensitivity of mean-variance-efficient portfolios to changes in asset means: Some analytical and computational results.. *he Review of Financial Studies*, p. 4: 315–342..

Brézillon, P., 1999. Context in problem solving: A survey. *The Knowledge Engineering Review*.

Bryce Goodman, S. F., 2016. European Union regulations on algorithmic decision-making and a "right to explanation".

Carmen Lacave, R. A. F. J. D., 2000. Graphical explanation in Bayesian Networks. *ISMDA Conference, Springer*.

Chandrasekaran, B. T. M. C. a. J. J. R., 1989. Explaining control strategies in problem solving. *IEEE Expert Vol 4 No 1 (Spring 1989) pp 9-24*.

Chen, L. a. B., 2019. Gmail Smart Compose.

Christian Meske, E. B., 2020. Using Explainable Artificial Intelligence to Increase Trust in Computer Vision.

CIFAR, 2017. CIFAR Pan-Canadian Artificial Intelligence Strategy.

Clanccy, W. J., 1981. The Epistemology of A Rule-Based Expert System: A Framework for Explanation.

Claudia Biancotti, P. C., 2018. Data superpowers in the age of AI: A research agenda.

Danding Wang, Q. Y. A. A. B. L., 2019. *Designing Theory-Driven User-Centric Explainable AI*. s.l., Conference: CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019).

Deloitte, 2016 (a). *The expansion of Robo Advisors in wealth management. White Paper.*, s.l.: s.n.

Deloitte, 2016 (b). *Robo Advisory in wealth management. White Paper.*, s.l.: s.n.

Ding, X. Y. Z. T. L. a. J. D., 2015. Deep learning for event-driven stock prediction.. *International Conference on Artificial Intelligence.*

Donna Lamberti, W. W., 1990. Intelligent interface design: an empirical assessment of knowledge presentation in expert systems. *ACM.*

Došilović, F. & B. M. & H. N., 2018. Explainable Artificial Intelligence: A Survey. *Conference: MIPRO 2018 - 41st International Convention Proceedings, At Opatija, Croatia.*

ETMoney [Mobile application software]. <https://www.etmoney.com/Edward>
Choi, M. T. B. J. A. K. A. S. W. F. S. J. S., 2016. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. *Accepted at Neural Information Processing Systems (NIPS) 2016.*

Finale Doshi-Velez, B. K., 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608.*

Flunkert, V. D. S. a. J. G., 2017. DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks.

FTAdvisor, 2018. *Robo advice suffers trust crisis.* [Online]
Available at: <https://www.ftadviser.com/investments/2018/06/19/robo-advice-suffers-trust-crisis/>
[Accessed May 2020].

G. Montavon, W. S. K.-R. M., 2017. G. Montavon, W. Samek, K.-R. Mller, Methods for interpreting and understanding deep neuralnetworks. *Digital Signal Processing 73 (2018) 1–15. doi:10.1016/j.dsp.2017.10.011..*

G. Octo Barnett, J. J. C. J. A. H. E. P. H., 1987. *DXplain: An Evolving Diagnostic Decision-Support System.* s.l., s.n.

Garcia, M., 2016. "Racist in the Machine".. *World Policy Journal.*, Volume 33 (4), p. 111–117..

Genest J, F. J. F. G. M. R., 2003. Recommendations for the management of dyslipidemia and the prevention of cardiovascular disease: summary of the 2003 update.. *CMAJ.*

Global Business Outlook, 2017. *Democratising finance: The digital wealth management revolution.* [Online]

Available at: <https://www.globalbusinessoutlook.com/democratising-finance-the-digital-wealth-management-revolution/>

[Accessed 2020].

Grégoire Montavon, A. B. L. S.-R. M., 2019. Layer-Wise Relevance Propagation: An Overview. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science, vol 11700. Springer, Cham.*

Guidotti, R. & M. A. & T. F. & P. D. & G. F., 2018. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys.*

Hai Shu, H. Z., 2019. Sensitivity Analysis of Deep Neural Networks. *AAAI Conference on Artificial Intelligence (2019), pp. 4943-4950.*

Hall, P., 2019. On the Art and Science of Explainable Machine Learning: Techniques, Recommendations, and Responsibilities.

Hess, T. F. M. a. C. D., 2009. Designing Interfaces with Social Presence: Using Vividness and Extraversion to Create Social Recommendation Agents..

Himabindu Lakkaraju, S. H. B. a. J. L., 2016. Interpretable decision sets: A joint framework for description and prediction.. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. pages 1675–1684..

Hoffman, R. R. M. S. T. K. G. & L. J., 2018. Metrics for Explainable AI: Challenges and Prospects.. pp. 1-50.

Holdren, F. G. B. L., 2016. Preparing for the Future of Artificial Intelligence. *Executive Office of the President*.

Hon, E., 2019. Decoding robo advisory for Indian investors. *MoneyControl*, 7 October.

IBM, n.d. *Building trust in AI*. [Online]

Available at: <https://www.ibm.com/watson/advantage-reports/future-of-artificial-intelligence/building-trust-in-ai.html>

Idzorek, T., 2005. A step-by-step guide to the Black–Litterman model: Incorporating user-specified confidence levels..

Insider, B., 2017. *Is Robo Investing better than traditional investing? See the pros and*. [Online]

Available at: <http://www.businessinsider.de/4-reasons-robo-investing-growing-2017-1%3fr%3dUS%26IR%3dT>

[Accessed 13 4 2020].

IOSCO, 2014. *Report on the IOSCO Social Media and Automation of Advice Tools Surveys*. [Online]

Available at: <https://www.iosco.org/library/pubdocs/pdf/IOSCOPD445.pdf>

[Accessed September 2019].

James Wexler, M. P. T. B. M. W. F. V. a. J. W., 2019. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and*.

James Wexler, M. P. T. B. M. W. F. V. J. W., 2019. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE VIS (VAST)*.

James Woodcock, Center for Drug Evaluation and Research, 2020. *FDA Continues to Support Transparency and Collaboration in Drug Approval Process as the Clinical Data Summary Pilot Concludes*. [Online]

Available at: <https://www.fda.gov/news-events/press-announcements/fda-continues-support-transparency-and-collaboration-drug-approval-process-clinical-data-summary>

[Accessed 2020].

- Jichen Zhu, A. L. , S. R. , R. B. G. M. Y., 2018. Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation. *2018 IEEE Conference on Computational Intelligence and Games (CIG)*.
- Jonathan Lazar, J. H. F. a. H. H., 2010. Research methods in human computer interaction.. *John Wiley & Sons*.
- Julia Angwin, J. L. S. M. a. L. K. P., 2016.
<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- Kass, R., 1991. Building a user model implicitly from a cooperative advisory dialog. *Springer*.
- Kenneth Holstein, J. W. V. H. D. I. M. D. H. W., 2019. Improving fairness in machine learning systems: What do industry practitioners need?. *ACM CHI Conference on Human Factors in Computing Systems (CHI 2019)*.
- Kim, T. W., 2018. Explainable artificial intelligence (XAI), the goodness criteria and the grasp-ability test. *arXiv:1810.09598*.
- Krishnan, S., Deo, S. & Sontakke, N., 2020. *Operationalizing algorithmic explainability in the context of risk profiling done by robo financial advisory apps*, s.l.: Data Governance Network.
- Laboure, M. & Braunstein, J., 2017. *Democratising finance: The digital wealth management revolution*. [Online]
Available at: <https://voxeu.org/article/digital-wealth-management-revolution>
[Accessed October 2019].
- Lacave, D., 2004. A review of explanation methods for heuristic expert systems. *Cambridge University Press*.
- Leilani H. Gilpin, D. B. B. Z. Y. A. B. M. S. a. L. K., 2019. Explaining Explanations: An Overview of Interpretability of Machine Learning.
- Limaye, V., 2019. The rise of small-town investors in Indian equity markets. *Economic Times*, 24 September.

- Lipton, Z., 2018. Mythos of model interpretability. *2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*, New York, NY.
- Lisa Anne Hendricks, Z. A. M. R. J. D. B. S. T. D., 2016. Generating Visual Explanations.
- Lombrozo, J. J. W. T., 2010. The Role of Explanation in Discovery and Generalization: Evidence From Category Learning. *Cognitive Science: A multidisciplinary journal*.
- Ludwig, L., 2020. *Investor Junkie*. [Online]
Available at: <https://investorjunkie.com/robo-advisors/cost-comparison/>
[Accessed 27 04 2020].
- Malin Eiband, H. S. M. B. J. F.-C. M. H. H. H., 2018. *Bringing Transparency Design into Practice*. s.l., s.n.
- Marco Tulio Ribeiro, S. S. C. G., 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv:1602.04938 [cs.LG]*.
- Marco Tulio Ribeiro, S. S. C. G., 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. s.l., s.n.
- Marco Tulio Ribeiro, S. S. C. G., 2018. Anchors: High-Precision Model-Agnostic Explanations. *AAAI*.
- Margaret Mitchell, S. W. A. Z. P. B. L. V. B. H. E. S. I. D. R. T. G., 2019. *Model Cards for Model Reporting*. s.l., s.n.
- Markoff, J., 2016. How Tech Giants Are Devising Real Ethics for Artificial Intelligence. *N.Y. TIMES*.
- Markowitz, H., 1959. Portfolio selection: Efficient diversification of investment.. *New York: Wiley*..
- Matt Thomas, Andy Masters, 2016. *Robo Advice: Revolutionizing the investment landscape*, United Kingdom: KPMG.
- Maurell, v. d. R., 2019. *Embracing Robo Advisory looks promising or the longitivity of Financial Advisors*, New York: s.n.

Mayank Bansal, A. K. A. O., 2018. ChauffeurNet: Learning to Drive by Imitating the Best and Synthesizing the Worst.

MEITY, G. o. I., 2019. Constitution of four Committees for promoting Artificial Intelligence (AI) initiatives and developing a policy framework.

Métayer, C. C. a. D. L., 2019. Understanding algorithmic decision making: Opportunities and challenges. *European Parliamentary Research Service*.

Michael R. Wick, W. B. T., 1992. Reconstructive expert system explanation. *ACM*.

Michaud, R., 1989. The Markowitz optimization enigma: Is optimized optimal?. *Financial Analysts Journal*, p. 45: 31–42..

Miller, T., 2018. *Explanation in Artificial Intelligence: Insights from social sciences*. s.l., arXiv:1706.07269v3 [cs.AI].

Miller, T., 2018. Explanation in Artificial Intelligence: Insights from Social Sciences. *arXiv:1706.07269v3 [cs.AI]* .

Mooney, M. B. a. R. J., 2005. Explaining Recommendations: Satisfaction vs. Promotion.

Morana, S. & G. U. & J. D. & G. C., 2020. The Effect of Anthropomorphism on Investment Decision-Making with Robo-Advisor Chatbots..

Moran, K., 2018. *Quantitative User-Research Methodologies: An Overview*.

[Online]

Available at: <https://www.nngroup.com/articles/quantitative-user-research-methods/>

[Accessed July 2020].

MutualFundIndia, n.d. *Mutual Funds India*. [Online]

Available at: <https://www.mutualfundindia.com/>

[Accessed 27 4 2020].

Nancy Staggers, A., 1993. Mental models: concepts for human-computer interaction research. *International Journal of Man-Machine Studies*.

Nielsen, J., 2000. *Why You Only Need to Test with 5 Users*. [Online]
Available at: <https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/>

[Accessed July 2020].

Nielson, 2006. *Quantitative Studies: How Many Users to Test?*. [Online]
Available at: <https://www.nngroup.com/articles/quantitative-studies-how-many-users/>

[Accessed July 2020].

Or Biran, C. V. C., 2017. *Explanation and Justification in Machine Learning*.

Patrick Hall, N. G., 2019. *An Introduction to Machine Learning Interpretability, Second Edition*. O'Reilly: s.n.

Patrick Hall, N. G. N. S., 2019. Proposed Guidelines for the Responsible Use of Explainable Machine Learning. *arXiv:1906.03533v3 [stat.ML]* .

PayTM money 2010 [Mobile application software].

<https://www.paytmoney.com/>

Pedro Antunes, V. H. S. F. O. a. J. A. P., 2012. Structuring dimensions for collaborative systems evaluation.. *ACM Computing Surveys*.

Pedro Saleiro, B. K. A. S. A. A. L. H. J. L. a. R. G., 2018. Aequitas: A Bias and Fairness Audit Toolkit.. *arXiv preprint arXiv:1811.05577*.

Philippe Bracke, A. D. C. J. a. S. S., 2019. Machine learning explainability in finance: an application to default risk analysis. *Staff working paper of the Bank of England*.

Qiu, L. a. B. I., 2009. "Evaluating Anthropomorphic Product Recommendation Agents: A Social Relationship Perspective to Designing Information Systems.. *Journal of Management and Information Systems*.

R.Swartout, W., 1983. XPLAIN: a system for creating and explaining expert consulting programs. *Elsevier*.

Robert Neches, R. F. T. F. T. G. R. P. T. S. a. W. R. S., 1991. Enabling Technology for Knowledge Sharing.

Roberto Confalonieri, T. W. T. R. B. F. M. d. P. M., 2019. TREPAN Reloaded: A Knowledge-driven Approach to explaining black box models..

Sai P. Selvaraj, M. V. S. R., 2018. Classifier Labels as Language Grounding for Explanations. *GCAI*.

Sandra Wachter, B. M. L. F., 2016. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation.

Sarah Holland, A. H. S. N. J. J. K. C., 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards.. *arXiv:1805.03677* .

Scott M. Lundberg, S.-I. L., 2017. A Unified Approach to Interpreting Model Predictions. *arXiv:1705.07874v2 [cs.AI]* .

Shane Mueller, R. R. H. W. C. A. E. G. K., 2019. Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications and Bibliography for Explainable AI. *DARPA XAI Program*.

Sharma, N. & A. L. M., 2010. Automated medical image segmentation techniques.. *Journal of medical physics*, Volume 35(1), 3–14. .

Shortliffe, B. G. B. a. E. H., 1984. Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project.

Simonite, T., 2017. *Machines Taught by Photos Learn a Sexist View of Women*.

[Online]

Available at: <https://www.wired.com/story/machines-taught-by-photos-learn-a-sexist-view-of-women/>

[Accessed 03 04 2020].

Singapore, A., 2018.

Singh, M. Y., 2019. Hierarchical interpretations for neural network predictions. *ICLR*.

Stanford, 2009. One Hundred Year Study on Artificial Intelligence (AI100).

Sven Coppers, J. V. d. B. K. L. , V. V., 2018. *Intelligo: An Intelligible Translation Environment*. s.l., 2018 CHI Conference.

Swartout, M., 1993. Explanation in Second Generation Expert Systems.

Tavaga 2018 [Mobile application software]. <https://tavaga.com/>

Timnit Gebru, J. M. B. V. J. W. V. H. W. H. D. I. K. C., 2018. Datasheets for Datasets. *arXiv:1803.09010*.

Tobias Plötz, S. R., 2018. Neural Nearest Neighbors Networks. *arXiv:1810.12575 [cs.CV]*.

William R. Swartout, J. D. M., 1988. *Explanation in Second Generation Expert Systems*. s.l., Springer, pp. pp 543-585.

Wojciech Samek, A. B. G. M. S. B. K.-R. M., 2015. Evaluating the visualization of what a Deep Neural Network has learned. *arXiv:1509.06321 [cs.CV]*.

ZHENG, X.-I., 2010. FinBrain: When Finance Meets AI 2.0. *Frontiers of Information Technology & Electronic Engineering*.

Conclusion

As AI increasingly becomes a part of our worlds and becomes especially noticeable in one of the most critical areas of our lives — our finances — the need for transparency grows too. Explainable AI addresses significant issues such as biases and transparency, which impact the users need to fully trust AI (Shin, 2021). While this seems conceptually intuitive, the process involves a number of “complex intellectual and scientific problems”. These problems have two solutions (technical and market suggestions) based on giving users a choice in who to trust (Wolfram, 2019). Thus, through these applications, it can be ensured that there will be better outcomes for everyone involved.

The European Commission High-Level Expert Group on AI had presented the Ethics Guidelines for Trustworthy Artificial Intelligence in 2019. These guidelines included some essential aspects of explainable artificial intelligence. They stated that regarding human agency and oversight, the decisions must be informed and that a provision for a human-in-the-loop oversight must be made. Regarding transparency, the explanations must be adapted to suit the concerned stakeholder. Further, users must be aware that they are interacting with an artificial intelligence system. Finally, when it comes to accountability of an artificial intelligence system, mechanisms for responsibility and accountability, auditability, assessment of algorithms, data, and design processes must be developed (Bussmann, Giudici, Marinelli, & Papenbrock, 2020).

The research has explored algorithmic explainability in the fintech sector, and found that the need for explainable AI in this sector is vital. It is needed to help companies understand how the AI algorithmic models can detect fraudulent transactions and minimise false fraud flags, thereby reducing expenses. The growth the global fintech market is experiencing, will inevitably lead to more financial services and fintech start-ups would leverage AI for various applications (Bussmann, Giudici, Marinelli, & Papenbrock, 2020). Thus, the need for transparency into such decisions made by these systems will only increase.

Through the three sections of research, significant conclusions were arrived upon, and the direction of future research was made clear. The first section looked into the practicality of adoption and limitations of explainability in India. Then, this research provided tools to generate explanations that satisfy regulatory and user-centric requirements. Finally, in the

third section, the grasp and acuity of these explanations were analysed to determine their effect on user trust in a decision-making system.

The first area of research focused on demystifying algorithmic and AI explainability for robo financial advisors in the Indian context. This research established that automated decision-making systems would inevitably be a part of emerging ecosystems within fintech in India. Further, as the number of people participating in these technology platforms increases, more innovation will take place, and efforts will be made to intelligently use artificial intelligence and machine learning to add value to entities.

After understanding the situation in various other countries and applying it to the Indian context, this research was able to draw out some astute observations. Through a series of interviews with regulators, individuals in the private sector, and users, it was found that people were pessimistic about the implementation of robust regulations needed to create a transparent and technologically advanced financial sector in the future. This was said to be due to private players distorting the market. It was also noted that there is a defined need for the maturity of responsible AI movement in India, which at the moment is slow-moving when compared to Western countries. Regarding the appropriate mode of regulation, interviewees were not in favour of a purely self-regulatory model since it would likely not be entirely effective in addressing the potential harms of automated decision-making systems. The interviewees were more inclined towards a co-regulatory model.

Further, considering the constraints of operating in the Indian context, not much could be said due to the limited number of investors. Some restrictions to AI adoption in financial services in India were identified to be the large number of customers that were new to the banking sector and without credit history; the language barriers to the deployment of conversational AI; and the lack of a dedicated citizens' forums to create awareness and air the voices of users. Finally, an inverse relationship between algorithmic explainability and algorithmic performance was found. Interviewees understood the challenges of explainable AI-driven decision making, but still cared for a certain level of transparency regarding those decisions even at the cost of performance.

Overall, the interviewees were found to have a cautiously optimistic view regarding the increased penetration of automated decision-making systems and their proper regulation.

This research also addressed the concerns surrounding the operationalising of algorithmic explainability in India from an operational perspective. Indian policymakers will likely consider approaches from Western countries first and then move on to launching consultations internally and coming up with an appropriate regulatory solution that would account for the unique challenges of this market. The factors that set India apart from other jurisdictions and consequently should inform policy formulations were found to be: size and population density, linguistic heterogeneity, and income and wealth inequality. It is recommended that an Indian algorithmic accountability regulator is set up.

Thus, this research was able to review the current explainability and regulatory landscape, along with existing limitations. Tools to satisfy the regulatory, as well as user-centric aspects of fintech applications were created. Finally, explainability was evaluated in the Indian context as compared to the international norms.

Through the second section of research, which looked into operationalising algorithmic explainability in the context of risk profiling done by robo financial advisory apps, it was found that the proposed method showed promise in its ability to enhance the technical capabilities of capital market regulators without the need for an in-house computer science expert. This was due to the fact that operationalising explainability in the case of robo-advisory risk profiling was done by the creation of a RegTech tool that would be used for several algorithms and use cases. The machine learning models used to recognise and reconstruct the three levels of explanations revealed the original risk profiling decision logic of the robo advisor. The three levels were: finding the importance of the inputs used in the risk profiling algorithm, inferring the relationships between said inputs and the assigned risk classes, and allowing regulators to explain decisions for any given user profile (as a method of spot-checking random data points). This is important as it provides regulators (who lack the technical knowledge to understand algorithmic decisions) a method to understand it. Further, the use of this by fintech regulators to audit algorithms and to check compliance with the regulations they are subject to is ensured.

Thus, this section was able to ascertain the aspects which make regulations successful and established the need for a dialogue between technologists. This can be achieved through regulatory sandboxes. Still, there is a need for work on the regulatory and technical fronts.

Specifically, there is a need to work with regulators to understand the graspability of various explanation methods and extend appropriate explanations to the user. For the technical front, there is a need for the initiation of the standardisation in addition to a robust documentation process for algorithms. This is extremely important to audit the system and maintain accountability.

The third section of research looked into algorithmic explainability in practice. This is done by evaluating the effectiveness of explanations in the context of robo-advisory apps. Through this research, a framework for user-centric and multi-perspective explanations of a robo financial advisory application-based on an interdisciplinary review of existing tools and methods was designed and implemented. The goals and information that users expected in explanations of complex decision support systems were identified, and the usability of the multi-perspective framework of explanations was examined. Further, the effect it has on users' trust was also examined.

It was found that users comprehended the explanations and had a positive response towards them (even the ones which were technical in nature). This showed a positive correlation between the users' trust and confidence in the system and the presence, comprehension of the explanation. This indicates that explanations of decision logic would benefit the user and developers of the system alike. This comprehension and trust were notably reduced between transparent white and opaque black box explanations of algorithms, and the difference of usability can be studied in future projects. Users are equipped to understand explanations of complex algorithms. Still, in general, there is a need for a more generalisable strategy of explanations that is complemented by cross-domain (legal, ethical, and policy) support. These results will be significant in helping policymakers and regulators understand user needs which will help in designing better policies around algorithmic explainability for robo financial advisors.

As a path forward, (Wolfram, 2019) reiterated the importance of finding solutions in which users don't have to trust the single AI of the automated content selection business. That is, the user must be able to pick their own brand of AI, which should be provided by a brand they trust. These third-party brands must be existing start-ups, media organisations, non-profits or something completely new — as long as they represent brands that users can trust. Such

a 'network' will eventually lead to something more significant than the existing monolith and give users the opportunity to freely pick from a whole market of AI (Wolfram, 2019).

The three papers in this series naturally come with their own set of research limitations. In the first paper, the fact that the Personal Data Protection Bill, 2019 is yet to become an Act had limited the extent to which the discussion around regulation and operating constraints could be specifically set in the Indian context. Since the upcoming law might significantly influence the processing of personal data in India, this limitation reduced the scope of the interviews with users, companies and regulators in this paper. For the second and third papers, there is a limitation in the *level of simplification* of a black box algorithm. As the papers explain, there is a trade-off between complexity, completeness, and accuracy of the system, and its explainability. This study is limited to developing a tool that can explain parametric and non-parametric models. Further, for the method developed in the second paper, which was also used in the third paper, the underlying assumptions on how a robo-advisory algorithm functions may pose as a limitation. However, the goal is not to develop a market-ready robo-advisory algorithm but rather to explain these assumptions transparently to users through explainable AI techniques. Finally, a limitation in the explanation is that we do not attempt to explain deeper layered models of neural networks, as the domain-specific requirements needed to develop those algorithms are beyond the scope of the study.

This series of papers also offers several takeaways for those specifically interested in the Indian domestic landscape involving ADS in financial services. Factors such as size and population density, linguistic heterogeneity and wealth inequality set India apart from the West, where the AI movement has already considerably matured. Other restrictions pertaining to the Indian public finance context, such as customers' lack of a credit history and familiarity with the banking sector, language barriers that come up in the deployment of conversational AI, and the lack of a relevant dedicated citizens' forums, act as additional unique operational constraints in India. But we now know that it is not a question of *if* ADS will play a significant role in India's financial services sector, but more a question of *when* that will happen, and that regulations are only likely to catch up at a later stage once the penetration has reached a significant level. At this early stage of the penetration of ADS in Indian financial services, both the industry and the users seem optimistic about the growth

of AI in this sector and pessimistic about the implementation of strict regulations needed to create a transparent and technologically advanced financial sector in the future, attributing the latter to the role of private players in distorting the market. Both the finance industry and users, however, agree on the need for the responsible AI movement in India to mature. They also seem to prefer a co-regulatory model of regulation to a self-regulatory model to adequately address potential threats and harms involved.

In the coming years, the need for AI will only accelerate. As AI revolutionises products or services across industries, the need to grasp how decisions are made and applied will increase and our ability to understand this logic must improve. Humans should be able to work alongside machines and must be able to trust that their systems are in place. In order to make that trust possible, explainability to understand these systems is needed.

Thus, companies must evaluate the trade-off between explainability and the performance of their algorithms in order to decide whether they should update their AI tools to remove the opaque black box in these algorithms. For a given algorithm which suits their needs, they must strive towards an improvement in explainability and in mitigating bias to improve outcomes for all (Infosys , 2019).

Appendix

Edited versions of two of the three chapters in this series have been published during the preparation of this dissertation. Details of the publications are as follows:

Chapter 2 — *Operationalising algorithmic explainability in the context of risk profiling done by robo financial advisory apps* — can be found at the following link

- <https://www.datagovernance.org/report/operationalizing-algorithmic-explainability-in-the-context-of-risk-profiling-done-by-robo-financial-advisory-apps>

and can be cited as

Krishnan S., Deo S., Sontakke N. (2020) *Operationalising algorithmic explainability in the context of risk profiling done by robo financial advisory apps*. Hertie School, Berlin, Germany.

Available at: <https://opus4.kobv.de/opus4-hsog/frontdoor/index/index/docId/3681>

- This paper has also won the “1st position for the [Best Paper Award](#)” at the SEBI-NISM Conference 2020.

Chapter 3 — *Algorithmic Explainability in Practice: Evaluating the effectiveness of explanations in the context of robo advisory apps* — is available in the following two locations

- https://link.springer.com/chapter/10.1007/978-3-030-78642-7_64

which can be cited as

Deo S., Sontakke N. (2021) User-Centric Explainability in Fintech Applications. In: Stephanidis C., Antona M., Ntoa S. (eds) *HCI International 2021 - Posters*. HCII 2021. Communications in Computer and Information Science, vol 1420. Springer, Cham. https://doi.org/10.1007/978-3-030-78642-7_64

- And <https://ieeexplore.ieee.org/abstract/document/9548021>

which can be cited as

Deo S., Sontakke N. (2021) Usability, User Comprehension, and Perceptions of Explanations for Complex Decision Support Systems in Finance: A Robo-Advisory Use Case. *Computer* 54(10):38–48. doi: 10.1109/MC.2021.3076851.

References for Introduction (pages 4-10)

- AXIS bank. (2020). Retrieved from https://www.axisbank.com/docs/default-source/press-releases/axis-bank-unveils-automated-voice-assistant-axaa.pdf?sfvrsn=7d937356_6
- Baer, D. (2019, November). *The 'Filter Bubble' Explains Why Trump Won and You Didn't See It Coming*. Retrieved October 2019, from The Cut: <https://www.thecut.com/2016/11/how-facebook-and-the-filter-bubble-pushed-trump-to-victory.html>
- CitiBank. (2018). Retrieved from <https://www.citibank.com/tts/about/press/2018/2018-1219.html>
- Costello, M., Hawdon, J., Ratliff, T., & Grantham, T. (2016, May). Who views online extremism? individual attributes leading to exposure. *Computers in Human Behavior*.
- Dastin, J. (2018, October 10). *Amazon scraps secret AI recruiting tool that showed bias against women*. (Reuters) Retrieved September 2019, from <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MKo8G>
- Datta, A., Tschantz, M. C., & Datta, A. (2015, February 18). Automated Experiments on Ad Privacy Settings. *Proceedings on Privacy Enhancing Technologies*, 92-112.
- Garcia, M. (2016). "Racist in the Machine". *World Policy Journal.*, 33 (4), 111–117.
- Guidotti, R. &. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*.
- HDFC Bank. (n.d.). Retrieved from <https://www.hdfcbank.com/personal/ways-to-bank/eva>
- Hon, E. (2019, October 7). Decoding robo advisory for Indian investors. *MoneyControl*.
- IBM. (n.d.). *Building trust in AI*. Retrieved from IBM What's next for AI: <https://www.ibm.com/watson/advantage-reports/future-of-artificial-intelligence/building-trust-in-ai.html>

Kari, P. (2019, October 25). *Healthcare algorithm used across America has dramatic racial biases*. Retrieved October 2019, from Guardian:
<https://www.theguardian.com/society/2019/oct/25/healthcare-algorithm-racial-biases-optum>

Krishnan, S., Deo, S., & Sontakke, N. (2020). *Operationalising algorithmic explainability in the context of risk profiling done by robo financial advisory apps*. Data Governance Network.

Limaye, V. (2019, September 24). The rise of small-town investors in Indian equity markets. *Economic Times*.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019, October 25). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*.

ProPublica. (2016, May 23). *Machine Bias There's software used across the country to predict future criminals. And it's biased against blacks*. Retrieved September 2019, from <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

Yudkowsky, E. (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk. *Global Catastrophic Risks*.

References for Conclusion (pages 207-210)

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020, April 24). Explainable AI in FIntech Risk Management. *Frontiers in Artificial Intelligence*, 3 (26), pp. 1-5.

Infosys . (2019). *Unlocking the Black Box with Explainable AI*. Infosys Knowledge Institute.

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, 146.

Wolfram, S. (2019, June 25). Testifying at the Senate about A.I.-Selected Content on the Internet.