氏　　　名	岡崎　孝太郎
学位(専攻分野)	博士(情報学)
学位記番号	総研大甲第 2457 号
学位授与の日付	2023 年 9 月 28 日
学位授与の要件	複合科学研究科　情報学専攻 学位規則第6条第1項該当
学位論文題目	Uncovering Dynamics and Viewers' Attitudes from Over-the-Top Media Service
論文審査委員	主　査　井上　克巳 情報学コース　教授 岡田　仁志 情報学コース　准教授 水野　貴之 情報学コース　准教授 市瀬　龍太郎 東京工業大学　工学院　教授 佐藤　真一 国立情報学研究所　教授

博士論文の要旨

氏　　名　岡崎孝太郎

論文題目　Uncovering Dynamics and Viewers' Attitudes from Over-the-Top Media Service

Recently, AI technologies have begun to penetrate widely and rapidly into our daily lives. However, they are black-boxed AIs because the behavior inside their models is too complex, and they cannot output predictions in a way that makes sense to humans. The explainability that the increasingly popular explainable AI(XAI) aims to achieve is still merely a mathematical summary of the internal states of black-box AI. This is nothing more than visualization of the processing 'reasons', limited only to the data science community. It is challenging for the current achievements of XAI research to ensure more advanced utilization of AI expected and the accompanying societal demands such as transparency, explanation accountability, and fairness.
The theme of explainability in human-AI systems is especially important in business environments, where there are large differences in information literacy, but where decisions must be made with high stakes. Thus, this study focuses on the domain of business decision-making. The subjects are decision-making groups at various levels within companies and organizations. The object is an observational space synthesized from endogenous in-company management resource activities and exogenous market environments.
The focus of the research is on a framework for developing XAI into Decision Intelligence(DI), and studies that have the potential to support techniques for transforming XAI "reasons" into DI "evidence" can be found in a number of other fields. Particularly noteworthy are resource rationality in cognitive science and predictive multiplicity demonstrated in data modeling research.
In cognitive science, human rationality is posited as the primary motivation for decision-making. This rationality is always incomplete and bounded. Therefore, social group decision-making based on empirical heuristics is required. The rationality that humans exhibit in cooperation with external knowledge, called cognitive resources, is called resource rationality. The research aims to achieve both of them by dividing the problem into two tasks, i.e., solving the problem and optimally allocating cognitive resources. In this study, we incorporate the concept of resource rationality into the design of human-AI systems for business decision making.
There are traditionally two cultures in mathematical modeling: data modeling and algorithmic modeling. The former hypothesizes interpretable structures in advance when designing a process to obtain output y from input x, and optimizes those

parameters. The latter considers the process's internals as invisible and algorithmically explores the relationship between x and y. As the perspective of XAI research is based on reflections on the excessive reliance on the latter, initiated by deep learning, the focus has been on as universal an explanatory methodology as possible, premising a trade-off between interpretability and accuracy. However, recently there have been frequent research results questioning the effectiveness of universal XAI technology in practical fields. Predictive multiplicity is the assertion that, contrary to the aforementioned trade-off, there exist several optimal models that achieve both accuracy and interpretability, according to pairs of domain-specific data characteristics and tasks. This research is grounded in the rationale of predictive multiplicity. Instead of aiming for global explainability, it focuses on learning strategies based on domain-specific data characteristics and tasks. It explores the interactions between business decision-making and AI systems in this context.

The question posed in this research is, how do social decision frame and AI systems interact through explainability? To address this issue, we focused our research on a framework to evolves XAI into DI, with business decision-making as the target domain. As a method for achieving explainability, we proposed the structuring of cognitive resources that promote domain-specific learning strategies, by leveraging resource rationality and predictive multiplicity. Specifically, we constructed and tested a group of systems that use viewing logs of Over-the-Top(OTT) media services as input and output a set of rules to support decision-making. All of these systems are commonly composed of task-solving flows at the multi-layered cognitive resource level. As a design method, we proposed explainable alignment. On top of this alignment, social decision frames and AI systems are placed as conceptual pairs according to the constraints of domain-specific data characteristics and the requirement of expected explainability.

The academic contribution of this thesis lies in the discovery of complementarity acting on both elements of this conceptual pair, and in its practical application. This complementarity can be defined as follows.

Let d be a domain-specific data characteristics constraint. A hypothetical space of explainability $E_d = (S_f, A_i)$ is a pair of sets of social frame S_f and a set of AI systems A_i such that E_d satisfy d. A complementary for a property p is a pair (S, A) such that $S \in S_f$, $A \in A_i$ and $S \otimes A$ satisfy p. S and A are complementary, that is, they integrate domain-specific explainability by the action of extrapolating relative perspectives from S to A and from A to S at the same time. In the model fusion, and in the choice of experimental methods, it is a necessary condition to be an element of this complementarity set. Thus, the conceptual pairs of social frames and AI models that fulfill this complementarity can be found at the level of each of the multi-layered cognitive resources. This complementarity allows us to recognize multiplicity of explanations by switching viewpoints, so to speak, like a trompe l'oeil, without any interface between the social frame and the AI model, and raises the explanatory potential of XAI to DI.

In this study, various methodologies are aligned based on this complementarity at different cognitive resource levels. Broadly, they can be classified into at least the following four types.

1. Isomorphism: $S \simeq A$. The symbol $\simeq$ denotes a relation that is “the same”. S is an isomorphism of A when objects S and A are considered structurally the same.
2. Allegory: $S \approx A$, where $\approx$ is a symbol indicating a “representing” or “symbolizing” relation. An object S is an allegory of A when it represents its own meaning while at the same time representing A, which is different from S.
3. Completion: $\varepsilon > 0, s \in S, a \in A, \|s - a\| < \varepsilon$. For any $\varepsilon > 0$, object S and A are complete if the above equation holds and the entire space is formed by S and A.
4. Projection: $S: A \to S(A)$. When object S assigns some output to object A, S is a function of A, projecting A to $S(A)$.

These complementarities in this study exhibit distinct characteristics under the synthesized requirements of domain-specific data characteristics and social explainability. Figure 1 shows their correspondence.

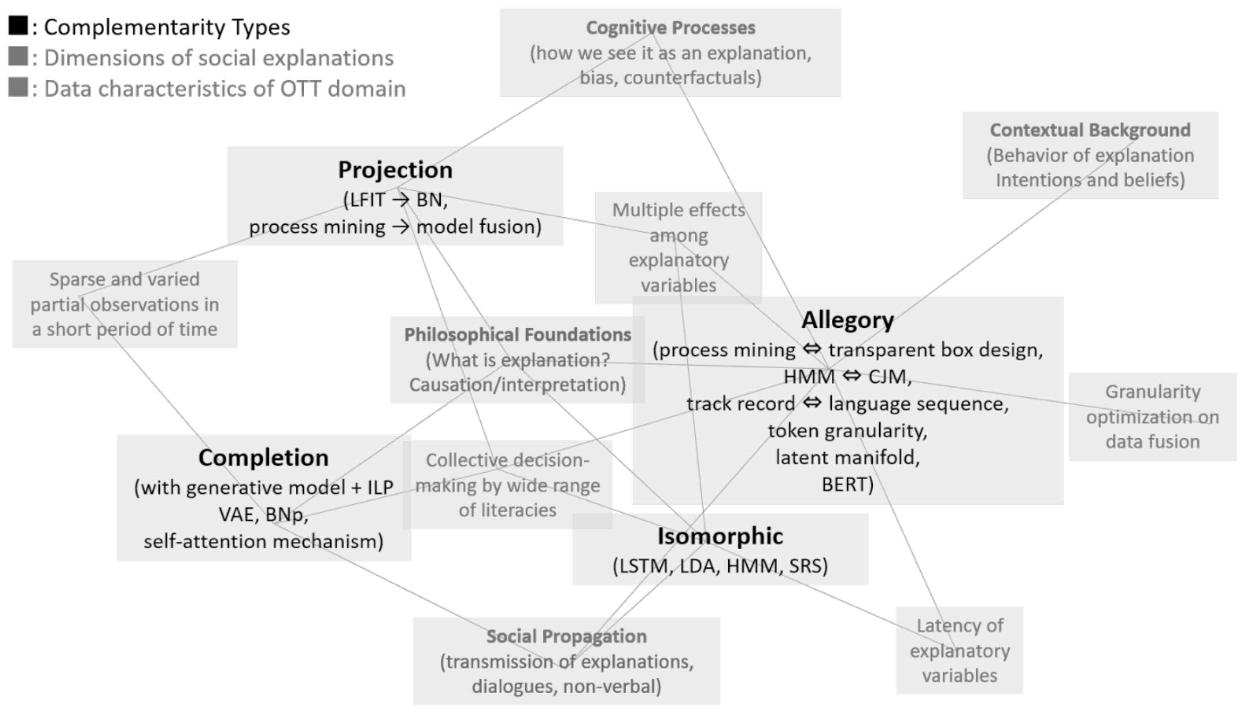

FIGURE 1: Complementarity types, dimensions of social explanations and data characteristics of OTT domain.

Assuming domain-specific data characteristics, choosing complementarity that matches the explainability required each time a social process is advanced, and deploying corresponding AI models accordingly - this is nothing but explainable alignment.

Figure 2 is a generalization of the explainable alignment in the OTT domain. At the center are pairs of transparent box designs and process mining, into which learners or

Managerial implications (Social decision-making frame)			$\mathbf{D}^u$: Observations updated		Theoretical implications (Explainable AI frame)
Domain specific data characteristics					
Generation of predictions for data assimilation	Conformance checking		$\mathcal{L}_b(\mathbf{D}^u), b(\mathbf{D}^u)$: Black box learner/predictor	Discretization preprocessing for data entry	Black box to ensure accuracy
Extracting latent attitudes of viewers Narrowing down attitudes to consider Visualization of the transaction process		Discovery	$\mathcal{L}_{c_a}(\mathbf{D}^a), c_a(\mathbf{D}^a)$: Comprehensive learner/classifier	Tuning of hyperparameters to optimize feature detection power	Static latent feature extraction behind explanatory variables Dimension reduction
Interpretation of black boxes from re-sampled focused data Visualization of the transaction process		Enhancement	$\mathcal{L}_{c_r}(\mathbf{D}^f), c_f(\mathbf{D}^f)$: Comprehensive learner/classifier		Visualization inside the black box as a surrogate model
			$\mathcal{L}_{c_p}(\hat{\mathbf{Y}})$, : Comprehensive learner	–	
Model credibility visualization			$accuracy(\mathbf{Y}, \mathbf{D}^u), fidelity(\hat{\mathbf{Y}}, \bar{\mathbf{Y}})$: Explainability Algorithmic framework	–	Distance measurement between ground truth and predictions Fidelity measurements for black box and surrogate models

FIGURE 2: Summary table for managerial and theoretical implication of algorithmic frame- work on explainable alignment.

explainers as instances can be fit. The far left and right ends are notations of positioning from the perspectives of social frames and AI systems, respectively.

Managerial implications (Social decision-making frame)			$\mathbf{D}^u$ Observations updated			Theoretical implications (Explainable AI frame)
1. Partial observations within short periods of time 2. Latent and confounding of explanatory variables 3. Appropriate granularity for data fusion 4. Process visualization						
			Chapter.3	Chapter.4	Chapter.5	
Generation of predictions for data assimilation	Conformance checking		LSTM (Recerrent Neural Model)	BERT (Transformer-type Neural Language Model)	NARM and STAMP (session-based recommender system)	Black box to ensure accuracy
Extracting latent static attitudes of viewers Narrowing down attitudes to consider Visualization of the transaction process		Discovery	LDA	Learned multi-head self-attention flow in BERT	Learned multi-head self-attention flow in NARM and STAMP	Static latent feature extraction behind explanatory variables Dimension reduction
Extracting latent dynamic attitudes of viewers Narrowing down attitudes to consider Visualization of the transaction process			VAE			Dynamic latent feature extraction Dimensional compression
Automated Ecological modeling from re-sampled focused data Visualization of the transaction process		Enhancement	HMM	PCFG (Probabilistic Context Free Grammar) tree Bayesian networks	–	Computing probabilistic automata Full observation pairing by stationary distribution
Logical rule outputs Output of an implication graph describing target features Visualization of the transaction process			BNp+ LFIT:GULA		LFIT:GULA	Rule learning from weighted interpretation transitions Implication graph generation
Model credibility visualization			$accuracy(\mathbf{Y}, \mathbf{D}^u)$, $fidelity(\hat{\mathbf{Y}}, \bar{\mathbf{Y}})$	Prediction of behavior by benchmark task (GLUE: SST-2)	Expert opinions(10ss)	Distance measurement between ground truth and predictions Fidelity measurements for black box and surrogate models

FIGURE 3: Summary table for managerial and theoretical implication for each chapters.

Figure 3 is arranged so that the methods employed in the experiments in Chapters 3, 4, and 5 can be compared and contrasted. In the black box AI section, LSTM was employed as an isomorphism, but it was later changed to a self-attention mechanism, which is both allegorical and complete, with an emphasis on accuracy reliability. In process mining, where discovery, namely the detection of latent variables, is prioritized, we mainly employed isomorphism as a highly interpretable complementarity due to the emphasis on explanatory power. Specifically, LDA, a hierarchical Bayesian model, and Session-based Recommender System (STAMP and NARM), which structurally incorporates two types of viewing attitudes into the modeling in advance. Since complementarity with high aggregation and interpretation power is suitable for enhancement, i.e., knowledge transformation of outputs in process mining, GULA, a generalization of Learning from Interpretation Transition(LFIT), is employed as a projection.

As described above, this study has shown that, practically speaking, a broader and more inclusive explainability can be achieved between social frames in the OTT domain and AI model fusion by applying appropriate complementarity at each stage of the process. Academically, we have shown that complementarity can be found as an interaction between the two via explainability, and that this complementarity can be developed in various ways according to the demands of domain-specificity. We claim that our study is original as a cross-disciplinary achievement on a cutting-edge issue of explainability.

The methodology proposed in this study shows strong potential to contribute universally to evidence-based decision-making based on large-scale data in all areas of the business domain. Although this research focused solely on the explainability of human–AI systems applicable to business, the experimental results have indeed opened up multiple interesting research perspectives. It is not just a framework perspective for developing XAI into DI. The viewpoints include the exploration of the nature of the manifold itself as a semantic space of various domains in the real world, as well as neuro-symbolic application research that realizes deductive inference through array computing on the distributed representation tensor obtained by induction on the real world. Future work will involve continuing to work on these related and cutting-edge re- search issues using real-world data, and to continue moving back and forth between practice and research.

（別紙１ Separate Sheet 1）

Results of the doctoral thesis defense

博士論文審査結果

Name in Full
氏 名　　岡崎　孝太郎

Title
論文題目　　Uncovering Dynamics and Viewers' Attitudes from Over-the-Top Media Service

本学位論文は、「オンライン型動画配信サービスにおける視聴動態と顧客態度の解明技術」（論文和題目）と題し、ビジネス領域に固有の特性をもつ行動データを入力として、ダイナミクスや潜在的な態度を論理プログラムの形式で抽出する方式において、ニューラル言語モデルと生成モデルおよび帰納論理プログラミングを経営工学フレームであるプロセスマイニング上で統合することにより、人工知能（AI）技術に詳しくない意思決定者にとっても透明性や公正性を有する頑健な説明を実現する新たな手法を確立した研究について述べている。本学位論文は英語で執筆されており全７章から構成されている。

第１章では、本研究で扱う固有領域であるオンラインコンテンツ配信サービスであるOTT サービス（Over-the-Top media service）とそのマーケティング技術、説明可能な人工知能（explainable AI; XAI）とその社会適用である決定知能（Decision Intelligence）について紹介し、高度な説明可能性へ向けた社会理論フレームと学習モデルとの相互作用について研究の背景が述べられている。マーケティング戦略の自動化のみならず、一般的に人間-AI システムにおける説明の追究は、タスクの解決方法を説明するアルゴリズムの問題であるとともに、人間の限界合理性にもとづく不完全さを常時補う認知資源の設計の問題でもある。本研究ではビジネス領域での人間-AI システムにおける説明について、社会理論フレームと AI モデルの概念対を用いて、精度と解釈性を両立するフレームワーク設計とその適用を提案している。

第２章では本論文で必要となる基礎知識として、ニューラル言語モデルや生成モデルおよび確率オートマトンの基礎、状態遷移を入力とし遷移規則を表現する論理プログラムを学習する「状態遷移からの学習（Learning from Interpretation Transition (LFIT)」方式、さらに本研究で用いている関連原理について説明している。

第３章は本研究の基本的アイデアとなる、カスタマージャーニーマップ作成のためのXAI モデルの基盤技術を提案しており、視聴データから LDA や LSTM などの機械学習技術を用いてユーザの視聴動態・顧客態度に関する状態遷移モデルを生成し、LFIT を用いてモデルから説明可能な規則を抽出するための一般的な方法について述べている。ここで提案する説明生成方式は「説明可能なアライメント」と呼ばれており、領域固有のデータ特性を制約とし、期待される説明可能性の要件に従って、社会理論フレームと AI モデルの相補的な対を用いてアルゴリズムが構築され、人間-AI システムにおける説明の多重性に対応している。ただしこの設定においては、行動系列に則した再帰的な学習の計算量的な問題や、規則を抽出する状態遷移が一次マルコフ性による記憶保持的問題を有していることが解析的に述べられている。

第４章では、第３章における問題の解決に向けて、行動系列学習にトランスフォーマー型の言語モデルを実装レベルで起用した新たなモデル融合を提案している。本手法では、行動系列の長期にわたる特徴量を保持することを目的とした双方向トランスフォーマーの利用や、潜在的な特徴量を蓄えたマルチヘッド・セルフアテンションフローからの状態遷移の直接出力、説明の方向に応じた確率文脈自由文法（PCFG）やベイジアンネットワークで表現されるオートマトンモデル生成、GLUE ベンチマークを使った行動予測等が含まれる。トランスフォーマーを使ったモデル融合の実現手法と各技術による規則抽出への影響に関する解析が述べられた後、スケーラビリティと説明可能性を検証するための実験を行い、各手法の有効性を詳細に分析している。

第５章ではトランスフォーマー型モデル融合の拡張として、長期的な嗜好性と短期的な駆動性という、視聴態度の二つの特質を識別するモデル構造を予め内蔵した系から、それぞれの特質を合成する規則を抽出するフレームワークを提案している。提案方式は、アテンション機構に基づくセッションベース推薦システム（SRS）を用いて実装されており、実験により各学習法の説明可能性を検証している。

第６章では人間-AI システムの説明可能性を含む関連研究について詳細にまとめており、第７章に本論文の成果をまとめ今後の課題について述べている。

公開発表会では博士論文の章立てに従って発表が行われ、その後に行われた論文審査会及び口述試験では、審査員からの質疑に対して適切に回答がなされた。質疑応答後に審査委員会では、出願者は情報学分野の十分な知識と研究能力を持つと認められ、博士研究は人間-AI システムの説明可能性の研究において十分なレベルの新規性を有しており、学術的に重要な貢献をなしていることが評価された。

以上を要するに本学位論文は、行動データから系のダイナミクスや潜在特徴を論理プログラムの形式で実現する際に、社会フレームと AI モデルの認識的な相補性を用いて、プロセス全体の精度と解釈性を両立するという画期的なフレームワークを提案し、可能な限りの効果効率性改善を試みることで、高い説明可能性を備えた、ビジネス領域での人間-AI システムの新たな可能性を示したものである。また、本学位論文の成果は、学術雑誌論文及びフルペーパー査読付き国際会議論文として発表され、学術的な貢献も認められる。以上の理由により、審査委員会は、本学位論文が学位の授与に値すると判断した。