Explainability in Mechanism Design: Recent Advances and the Road Ahead

Sharadhi Alape Suryanarayana[1,2], David Sarne[1], and Sarit Kraus[1]

[1] Department of Computer Science, Bar-Ilan University, Israel
[2] Centre for Ubiquitous Computing, University of Oulu, Finland
sharadhi.as@gmail.com, david.sarne@biu.ac.il, sarit@cs.biu.ac.il

Abstract. Designing and implementing explainable systems is seen as the next step towards increasing user trust in, acceptance of and reliance on Artificial Intelligence (AI) systems. While explaining choices made by black-box algorithms such as machine learning and deep learning has occupied most of the limelight, systems that attempt to explain decisions (even simple ones) in the context of social choice are steadily catching up. In this paper, we provide a comprehensive survey of explainability in mechanism design, a domain characterized by economically motivated agents and often having no single choice that maximizes all individual utility functions. We discuss the main properties and goals of explainability in mechanism design, distinguishing them from those of Explainable AI in general. This discussion is followed by a thorough review of the challenges one may face when working on Explainable Mechanism Design and propose a few solution concepts to those.

Keywords: Explainability · Mechanism Design · Justification.

1 Introduction

Intelligent systems and automated decision-making are replacing and enhancing human decision-making nowadays to the extent that people are increasingly reliant on them [42]. Despite the increased presence of such systems, people are not often aware that they are interacting with an AI-based system. Recognizing the need for transparency in this evolving policy and technology ecosystem, the ACM U.S. Public Policy Council (USACM) and ACM Europe Council Policy Committee (EUACM) codified a set of principles such as awareness, explainability, accountability, validation and testing to address this [20]. Among which, *Explainability*, which could be understood as a description in some form of the functioning of the system, has gained immense traction in the recent past.

Due to their opacity, domains with black-box algorithms like machine learning and deep learning have been extensively researched in the context of explainability. However, the need for explainability goes far beyond black-box algorithms. For example in various multi-agent systems (MAS), where agents are self-interested, commonly arises a need to aggregate private preferences such as availability, budget constraints and geographical location of several agents into

a collective decision in a socially desirable way. Mechanism design, an important tool in economics and computer science, is one such research topic which is concerned with the development of a mechanism that takes into consideration the preferences of selfish and intelligent agents exhibiting strategic behavior while adhering to norms such as envy-freeness, budget-balancing and pareto-optimality [57]. The applications of mechanism design can be found in various real-world and, in many cases, high-impact applications such as elections, rent division, resource allocation, and stable matching [28,57].

Regardless of their extensive usage in the real-world, there is a renewed interest in designing and analyzing mechanisms to align with human values. This includes re-designing existing mechanisms to accommodate human preferences [18], viewing existing practices for inclusive housing allocation from a game-theoretic perspective [8], empirical studies on human behavior [66], using the insights from empirical analyses to re-frame a mechanism [52], and devising algorithms to justify the decision of a mechanism [9]. Explaining the results to human participants is a natural and complementary extension to the pursuit of designing mechanisms that are more "understandable" to humans.

Nevertheless, the road to Explainable Mechanism Design systems is replete with its own share of hurdles. One key element in providing an explanation in such domains is the goal of the explanation and the measures of its success—whereas with a single user the system's goal is known, hence the explanation aims to improve her recognition of the optimality of the choice made, in settings of mechanism design, a user does not always know the system's goals since they may depend on other agents' preferences. This focus on preference aggregation of multiple agents, often associated with conflicting goals, may lead to a blatant compromise of the preferences of some of them. The explanations should therefore aim to increase user satisfaction by taking into account the system's decision, the user's and the other agents' preferences, the environment settings, and properties such as fairness, envy and privacy [29]. In addition to the above intricacies, the presence of domains such as voting, scheduling and resource allocation in popular culture, without the necessity to be theoretically aware, has led to people forming their own irrational opinions which the explanations have to uproot [16,38,67].

We also note that even cases where social choice is merely a particular step in a multi-stage decision making process carried out by an agent making decisions on behalf of humans, call for Explainable Mechanism Design. This can hold even when the use of mechanism design is not explicit. Examples of such settings include algorithmic hiring and virtual democracy (an approach to automated decision making) which is used in autonomous vehicles and kidney exchanges to automate moral decisions, and recommendation systems to allocate food donations to recepient organizations [18,36,49,62].

In this survey, we provide a comprehensive summary of the various threads of explainability in mechanism design. We note that while the broad theme of explainability translates to the same meaning with respect to both machine learning and mechanism design, there are a few differences between the premises

of the two fields in terms of what leads up to generating explanations. In particular, we first provide a comparison between Explainable Mechanism Design and Explainable AI (XAI) with respect to the taxonomy, the purpose of explanations and who the explainees are. We then outline the methods of generating explanations in mechanism design. Finally, we elaborate on the challenges of conducting laboratory experiments on explainability in mechanism design and shed light on solution concepts combining insights from XAI and behavioral studies.

2 Mechanism Design

Various definitions for mechanism design have been suggested over the years [47,53]. Essentially, a mechanism can be seen as a "communication system" where the participants send messages to each other and/or to a "message center" and every collection of messages is assigned an outcome based on a pre-specified rule [25]. These messages are characterized by private information such as utility from an allocation (in rent division), preference over a set of candidates (in social choice theory), or willingness to pay for a good (in auctions). Thus, the mechanism is analogous to a machine that collects, processes and aggregates the private information of several agents in order to reach a desirable social outcome. In most cases the agents are self-interested and rational, and care only about maximizing their private utility with no guarantee that they will tell the truth. Therefore, a desired property of a mechanism is that the agents have no strategic incentive to deviate from truth-telling [33]. A mechanism satisfying this property is considered *incentive-compatible*, as every participating agent achieves the best outcome by reporting her true preferences [33].

Since incentive-compatibility is a desired feature, most mechanisms are designed to be incentive-compatible. Hence, even though truth-telling might fetch the least utility, say not winning the Vickrey-Clarke-Groves (VCG) auction [33], it is the best response for an agent in most mechanisms. However, as we observe in section 7, due to repeated interactions, humans may be prone to their own biases. Explanations could be a tool to mitigate user biases as well.

The two main branches of designing mechanisms are the axiomatic branch and the Bayesian branch [14]. In the axiomatic branch, the solution is supposed to satisfy a set of desired properties called axioms. Axioms are normative elements designed to conceptualize notions of reason such as fairness, justice and efficiency. Examples of such axioms include envy-freeness and pareto-optimality. In the Bayesian branch the solution achieves an optimal value for a given objective function such as expected revenue or projected loss.

3 Motivations for Explainable Mechanism Design

Consider the example of a rent division problem with 3 housemates, 3 rooms with a total rent of \$3. Housemate i values room i at \$3 and the other two rooms at \$0. One possible solution to this problem is to assign room 1 to housemate 1 at \$3 and rooms 2 and 3 to housemates 2 and 3 for free. Even though from

an inter-personal perspective, this solution seems blatantly unfair to housemate 1, it is an envy-free solution, hence theoretically acceptable. Housemate 1 is indifferent between the three rooms, given their cost, while housemates 2 and 3 are overjoyed [19]. While this is an overblown depiction of the problem with a rent division setting which has seen an immense improvement in the solutions proposed over the years [19], it is sufficient to illustrate the complexity with respect to devising explanations in mechanism design settings.

As explained in the former section, solutions in mechanism design are obtained by aggregating the preferences of several agents. The nature of the problem necessitates a "social approach" where the solution is expected to aggregate said preferences in some acceptable manner. This can be achieved by mandating the solution to adhere to certain desirable criteria that are egalitarian in nature, maximizing a criterion of social welfare or using any other method that appreciates the social nature of the problem. The need to balance the preferences of several agents, which could be conflicting in nature might result in the solution not being in favor of a few of them. Multi-user Privacy Conflict due to varying privacy preferences of owners of shared content [43], multi-attribute settings such as *team formation* where the solution might not adhere to the preferences of all of the agents with respect to every attribute [21] and settings such as the classical "glove game" where the solution is non-intuitive yet theoretically sound [48] are other real-life examples of mechanism design that necessitate a nuanced approach to obtain the solution as well as to devise explanations.

In addition to the social nature of the problem, solutions in mechanism design face two hurdles. First, the issue of what is socially acceptable can vary according to perception, context and domain which can result in multiple solutions for the same problem. For example, in social choice theory, there are multiple voting rules due to the absence of a unique voting rule that satisfies Arrow's mandatory principles of fairness [5]. The second problem which is a consequence of the first, is that it is easy for the user to challenge the solution. This requires the explanations to not only elaborate on how good the solution is but also how problematic another solution is.

This is in sharp contrast with other domains such as machine learning, planning and recommender systems where a solution, good or bad, results from a definite and often complicated algorithm which needs to be broken down for user understanding [15,42,69]. While solutions in other settings can also rely on metrics such as accuracy, the absence of a unique algorithm in mechanism design settings makes this problem hard as well. Hence, instead of arguing on the theoretical accuracy of the decision or how it results from a particular algorithm which is a common trait in the aforementioned domains, the focus of explanations in mechanism design should be on arguing how the decision is "good" in its social context and how the preferences of the agents have led to the solution.

4 Explainable Mechanism Design versus XAI

While both Explainable Mechanism Design and XAI aim to explain certain decisions made by the system, in retrospect, there are several factors differentiating the two. Owing to the differences in the domains and the solution concepts, there is an obvious difference in the nature of explanations offered in Explainable Mechanism Design and XAI. One way to reason about the differences in the nature of the explanation to be offered, is by considering the differences in the taxonomy to be used, the explainee and the goals of the explanations. Hence, we focus our comparison in these aspects. Since XAI has become a well established research area, whereas Explainable Mechanism Design is a newly emerging field, we use the first to lead the discussion, mapping and contrasting equivalent notions of Explainability in Mechanism Design accordingly.

4.1 Taxonomy

The common types of Explanations that are found in Explainability Studies are *Explanations*, *Justifications* and *Interpretations*. With respect to mechanism design, the most relevant and consequently most researched capability is Justifiability. *Justifications* deal with explaining the system's decision in terms of acceptable societal norms [32]. These norms (i.e. axioms that formalize desirable social concepts such as fairness, justice and efficiency) are the foundation of many mechanisms. Hence they are natural and, in many cases, effective contenders for explanations [9,48,57].

While there is no consensus on what constitutes an *Explanation* in mechanism design, we adopt the idea proposed by Langley [32], who states that *an intelligent system exhibits explainable agency if it can provide, on request, the reasons for its activities*. Still, the extent of research dedicated to this capability is somehow limited (see Table 1 later on). As for *Interpretability*, this notion is completely absent from Explainable Mechanism Design.

For example, in the domain of fair division, explaining the decision by highlighting how the decision is envy-free is a *Justification* [35]. However, comparing the maximin solution (the solution that maximizes the minimum utility for every player thus resulting in the least disparity) to an arbitrary envy-free solution to demonstrate the superiority of the former solution amounts to an *Explanation* [19].

However, in XAI, the relationship of Explanations and Justifications to the algorithm are reversed. Here, *Explanations* act as an accurate proxy of the model while still being understandable to the human users [4] and *Justifications* defend the decision of the algorithm by explaining why it is a good one without necessarily focusing on how the decision was made [59]. While nearly absent from *Explainable Mechanism Design*, Interpretability which aims to enhance user understanding and comprehension of the model's decision-making process and predictions [42] through *Interpretations* [59] is rigourously researched in XAI. It is also interesting to note that *Justifiability* is the most researched capability in *Explainable Mechanism Design* since there is a need to show how the decision

abides by desirable social norms while the need to explain the functioning of black-box nature of the algorithms has led to *Explainability* and *Interpretability* enjoying the most attention from the XAI research community.

4.2 Explanation Purpose

Explanations may be provided for various purposes and goals, with partial equivalences between Explainable Mechanism Design and XAI.

Appreciating System Decision-Making. In Explainable Mechanism Design, an appreciation of the system's decision translates to understanding how the preferences of the different agents are combined to obtain a collective decision [9]. This is akin to understanding how different features contribute to the output in a machine learning model [31], which is the focus of XAI. Still, as explained in Section 3, the task of increasing user appreciation of the system's decision making is more challenging in mechanism-design settings, as decisions need to be explained in their wider social context and are often plagued with impossibility theorems [5,60].

Improving User Trust and Reliability. With the decision depending on the preferences of several agents, dissatisfaction is an inevitable evil in mechanism design which might affect the user's trust and reliability. The explanations presented thus need to argue about the legitimacy of the decision and how, even if the decision is unfair to a particular agent, the mechanism as a whole has adhered to mandatory principles of fairness and the dissatisfied agent needs to make peace with it [29]. In fact, as demonstrated in the work of Suryanarayana et al. [65], it is often the explanations provided to those participants for whom the winning candidate is the least preferred that are the most impactful. In XAI, unfavorable situations are also present (e.g., rejection of a loan application), however the prospect of improving the odds of the decision being in her favor through *Counterfactual Explanations* exists [44]. This, unfortunately, cannot be achieved in *Explainable Mechanism Design*.

Ensuring Accuracy of Decision-making. In mechanism design, explanations can serve as a tool for the verification of results in order to ensure that the decision was made under a set of rules consistently applied in each setting [7]. This somehow resembles the use of XAI as a tool for bias mitigation and fairness assessment [42] in cases where datasets are potentially biased and decision-making might be discriminatory.

4.3 The Explainee

The nature and mode of the explanation to be offered depend on who the explainee is and the purpose of providing explanations to her [59]. One of the main recipients of explanations in *Explainable Mechanism Design* are the end users or

participants as they are the ones affected by the decision made.[3] Unlike end users in XAI who are typically passive (as the system is making the decision for them, based on their data) users receiving explanations in mechanism design play an active role in the collective decision process taking place, by reporting their preferences. Examples of such explainees include a researcher whose grant proposal was rejected [7], one among many roommates who is assigned a particular room and rent based on her actively reported (and hopefully true) preferences [19] or, a user in a hybrid domain like algorithmic hiring where a social choice function is applied in some part of the overall algorithm [62].

Two classes of explainees in Explainable Mechanism Design enjoy some similarity with those of XAI. The first one is the *Decision Maker* who does not need to be an expert in mechanism design but has to have relevant knowledge of the domain in order to make informed decisions. This could include the employees of a refugee resettlement agency who need to be able to override the decision proposed by the algorithm [2] or the members of a committee who cannot decide on a specific voting rule and base the election on a set of desired properties (axioms) [9]. The Decision Makers are similar to *Data Experts* in XAI who use explanations to visualize, inspect, tune and select models [42]. The second class is that of an *External Entity*, similar to its namesake in XAI [59], who is someone not directly interacting with the system, say an auditor who needs to ensure that the decisions made adhere to a set of rules and that there is no violation [7]. In both classes of explainees, the requirements of the explanations to be produced in XAI and Explainable Mechanism Design overlap.

5 Explanation Concepts

As with XAI [42], Explainable AI Planning [15] and Explainable Recommendation [69], we provide a brief overview of the theoretical aspects (natural contenders) and behavioral aspects (necessary for human comprehension) of the explanations available in literature. Table 1 depicts a breakdown of the surveyed papers with respect to the different concepts and the evaluation methods discussed in the next section.

5.1 Norms versus Attributes

As mentioned earlier, mechanism design has two defining characteristics - the private information (such as preference and cost) of the agents participating in it and the requirement for the solution to recognize its social nature. Both of these can be used to devise explanations. Norms that formalize the desirable social traits are the foundation of solutions in mechanism design and can hence be used to extol its virtues. Attributes, on the other hand, quantify the stake a given agent has in the mechanism. Explanations that relate the solution to an

[3] In many cases the need to provide users with proper explanations is dictated by the regulator, e.g., in the case of GDPR guidelines [29].

Table 1. Surveyed literature organized by Explanation Concepts – Normative Characterization (NO), Attributive (AT), Contrastive (CO), Argumentative (AR), Visualization (VI) and Evaluation Methods – Theoretical Properties (TP), Computational Complexity (CC), Empirical Analysis (EA) and User Studies (US).

Explanation Concepts and Evaluation Methods										
Work	Setting	NO	AT	CO	AR	VI	TP	CC	EA	US
Ahani et al. [2]	Refugee Resettlement		✓	✓		✓				
Belahcene et al. [7]	Approval Sorting	✓		✓	✓		✓			
Boixel and Endriss [9]	Voting	✓					✓		✓	
Boixel et al. [10]	Voting	✓					✓			
Boixel and de Haan [12]	Voting	✓						✓		
Cailloux and Endriss [13]	Voting	✓			✓		✓			
Gal et al. [19]	Rent Division	✓	✓	✓		✓				✓
Georgara et al. [21]	Team Formation		✓	✓					✓	
Kirsten and Cailloux [26]	Voting	✓		✓	✓				✓	
Knapp [27]	Matching Theory	✓					✓		✓	
Lee et al. [35]	Rent Division	✓	✓	✓		✓				✓
Mosca and Such [43]	Multi-User Privacy Conflict	✓	✓	✓	✓					✓
Nardi et al. [46]	Voting	✓					✓		✓	
Nizri et al. [48]	Payoff Allocation	✓								✓
Peters et al. [54]	Voting		✓				✓	✓	✓	
Peters et al. [55]	Voting	✓					✓		✓	
Pozanco et al. [56]	Scheduling		✓	✓						✓
Suryanarayana et al. [65]	Voting	✓		✓						✓
Zahedi et al. [68]	Task Allocation		✓	✓	✓		✓			✓

agent's individual stakes can be effective in helping her appreciate the impact of the solution from a selfish perspective and thus convince her. For example, convincing a housemate in a rent division setting that the decision is envy-free amounts to a normatively characterized explanation [35] while the comparison of the maximin (the solution that maximizes the minimum utility for every player thus resulting in the least disparity) solution to an arbitrary envy-free solution to demonstrate the lower disparity achieved by the former solution is an attributive explanation [19]. In the following paragraphs, we elaborate on diverse settings where both norms and attributes have been used to devise explanations.

Normative Characterization. Formally in mechanism design, axioms are used to capture the social norms that the solution is expected to adhere to. Procaccia [57] advocates for the use of axioms to not only be used for designing a mechanism but also to justify its solutions with an example of the not-for-profit website *Spliddit* [23]. Justifying an outcome using a set of agreed upon axioms, without having to depend on a particular rule has found a special appeal in the domain of social choice theory, where no unique outcome can be obtained while following fair voting procedures [5].

In social choice theory, given a voting profile, Cailloux and Endriss [13] developed a logic-based language to construct arguments for and against specific outcomes. Using the elements of the proposed language, an algorithm to justify the Borda outcome given a voting profile was developed. Building on said approach, Boixel and Endriss [9] developed a formal notion of justification based on the definition of Langley [32] and an algorithm based on constraint programming to compute the justifications using any set of axioms. To counter the computational complexity of the aforementioned algorithm [12], Nardi et al. [46] used a combination of instance graphs and state-of-the-art SAT solvers to design an algorithm that can provide viable justifications. To enhance the readability of the justifications using the axiomatic approach, Boixel et al. [10] developed a tableau-based calculus. Using a combination of SAT solving and Answer Set Programming to implement the calculus, the authors provide an insight into how the justifications look.

The evolution from Boixel and Endriss [9] to Boixel et al. [10] helps visualize the transformation from a non-automated procedure to an automated procedure, from unstructured justifications to structured justifications and, from manual post processing to obtain the justifications to tableau-based rendering of the justifications for enhanced readability. A demonstration summarizing the application of the aforementioned techniques proposed [9,10,45,46] to find justifications given a normative basis can be found in Boixel et al. [11]. Furthermore, the approach used by Boixel and Endriss [9] has been extended to matching theory [27] where an algorithm is designed to justify outcomes that are of interest to a given agent (local outcomes) instead of the whole outcome.

The axiomatic characterization is also used to derive justifications for the results of approval voting [55] and non-compensatory approval sorting [7]. In the broader sense, non-compensatory approval sorting and voting are concerned with aggregating collective information into a single decision. The reviewed literature on justifying the results of a voting mechanism reveal all of the preference ballots. However, Belahcene et al. [7] show that the classification based on the binary judgements of the participants is compliant with the decision making process by revealing minimal information that is backed by theoretical properties.

One of the key elements used in some of the papers is Automated Reasoning (AR) using SAT or SMT solvers [7,10,45]. This combination of AR with social choice theory can be used to identify if a voting rule satisfies a particular axiom (thus arguing against it) [26] as well as verifying the correctness of the system output [12].

User studies to test axiom-based explanations were also successful in increasing satisfaction. Suryanarayana et al. [65] tested explanations based on features constructed from axioms in the domain of ranked-choice voting while Nizri et al. [48] used the axiomatic characterization of Shapley value [64] to come up with explanations in the domain of fair division. The research carried out by Nizri et al. [48] is significant in two ways. First, the solution that is being justified, i.e. Shapley value, satisfies all of the desired properties of fair division [24]. Second, the axiomatic characterization is not only used to justify the solution but also

to come up with the algorithm to generate explanations. The authors decompose the coalitional game into sub-games and generate explanations for each of these games based on the additivity axiom which states that the sum of Shapley allocations in each sub-game is equal to the Shapley allocation in the original game. The explanations were successful in convincing the participants that the allocation was fair.

An exception in terms of the norms used can be found in the work of Mosca and Such [43] in the domain of Multiuser Privacy who propose an explainable agent called ELVIRA that collaborates with other ELVIRA agents to identify the optimal sharing policy for shared content. Here, instead of axioms, the authors use a socio-cultural theory of human values by Schwartz [63] known as the *theory of basic values*. The explanations however, are based on both values and individual attributes, i.e. the privacy preferences of the participants.

Attributive Explanation. If norms capture societal acceptance, attributes quantify personal interests. Hence, explanations that relate to the individual attributes of the participants have also been fruitful in increasing participant satisfaction

Zahedi et al. [68] compare the *cost* of a proposed allocation to the cost of the counterfactual allocation proposed by the participant. Ahani et al. [2] depict the change in *employment score* if the refugee allocation proposed by the algorithm needs to be changed. A tangential direction termed *priceability* where voters spend money on buying candidates which forms an intuitive explanation for the committee selected in approval-based committee elections was proposed by Peters et al. [54]. Explaining outcomes based on individual attributes enables the comparison of solutions that are equally good in terms of theoretical requirements. For example, Gal et al. [19] explain their optimal rent division solution by comparing it to another envy-free rent division.

Several explanation generation methods are procedure-agnostic i.e., do not focus on the procedure that leads to the outcome. Here, a framework is developed to encode the different facets of the problems such as explanations, queries and constraints. Notable examples include encoding the Justification Generation Problem for collective decisions into a Constraint Network [9], developing a generic procedure for providing justifications for Team Formation Algorithm (TFA) while keeping the TFA intact [21] and using Mixed-Integer Linear Programming (MILP) to explain why the preferences of a participants were not satisfied [56] while designing a preference-driven schedule. While using such frameworks simplifies the process of finding an explanation, adequate care needs to be taken to convert the explanations provided by the system into a readable form. One solution to this problem is using explanation templates [21,56].

5.2 Catering to the Human Mind

Understanding how human-beings explain and respond to explanations can reveal important insights into how explanations of a system can be presented. Two such behavioral modes of explanations that are considered effective and

that have found applications in mechanism design are *Contrastive* explanations and *Argumentative* explanations [41]. Given the fact that mechanism design settings are social in nature, it is imperative that the behavioral nature of the explanations are attended to. Most of the papers that use a behavioral element in their explanations incorporate an element of social-interaction [41], which is necessary for a layman to comprehend the functioning of a complicated AI-based system.

As mentioned earlier, mechanism design settings suffer from the issues of familiarity, non-uniqueness of the solution and the solution not being in favor of the participants. This provides the perfect ground for the participants to challenge the solution. Identifying this, there is a great deal of interest in devising *contrastive* explanations that provide reasons for why a particular event did not occur as opposed to why a particular event did [41]. From Table 1, we can see that in nearly all of the cases where the explanations are based on the individual attributes of the agents, they are contrastive. In these scenarios, contrastive explanations can help the participant compare the difference in her utility across different solutions and thus appreciate the decision better.

In Suryanarayana et al [65], a contrastive explanation comparing the winning candidate to the participant's most preferred candidate were found to increase user satisfaction and acceptance the most when the winning candidate was the least preferred option of the participant. Similarly, in Mosca and Such [43], contrastive explanations were found to be more appealing than general descriptive explanations when the recommended solution was different from the participant's preference. Contrastive explanations are also especially useful in multi-attribute/multi-preference settings where the outcome may not align with all of the preferences of any participant. Other notable studies that uses the contrastive approach are of Georgara et al. [21] that provide explanations for both collaboration queries (questioning team formation) and assignment queries (challenging the assignment of teams and individuals to tasks) at individual, local and global levels, and the work of Pozanco et al. [56] which provides contrastive explanations regarding the unsatisfied preferences of the participants while ensuring that the explanation is relevant to the participant.

As far as argumentation is concerned, the presence of umpteen conflicting axioms is an encouraging premise to build an argumentation framework. This is demonstrated by Cailloux and Endriss [13] who developed a formal framework for presenting arguments favoring a particular outcome. Zahedi et al. [68] present the case for the a suggested task allocation by demonstrating how a negotiation based on the counterfactual task allocation proposed by the participant can lead to a higher cost. Mosca and Such [43] base their explanation on the argumentation scheme used to obtain the optimal solution. Both in Zahedi et al. [68] and Mosca and Such [43] argumentation is used for devising the optimal solution which was then organically extended to generating explanations in favor of the outcome.

Visualization is a tool that is often viewed as a less technical means of conveying complex theoretical concepts [6]. Human-in-the-loop systems are a natural extension to mechanism design that caters to capturing reality better. Here, the

algorithms have the capacity to process large volumes of data while expert insights are required to handle the inherent uncertainty of the real world. Hence, in addition to enabling easier comparisons [19], visual tools can also be used to support human decision-making.

Notable illustrations of visualization can be found in the case of the resettlement agency *HIAS* that is involved in resettlement of refugees into communities in the USA. The matching software *Annie*$^{\text{TM}}$ *MOORE* enables the employees to override the proposed allocation by revealing the updated statistics so that no change will have a grievous impact [2].

Explaining the outcome through effective visualizations can aid in enhancing the appreciation of fairness, a theoretical notion that is the bread and butter of mechanism design, as was observed by Gal et al. [19]. While Ahani et al. [2] capture practical elements such as indivisible families of refugees, batching and, an unknown number of refugee arrivals in the context of refugee resettlement, Gal et al. [19] provide the *fairest* division of rent subject to envy-freeness. In both of these cases, the practical relevance and efficacy of the proposed algorithm is demonstrated with the help of explanations. Hence, user studies with explanations can be seen as a complementary extension to establishing the superiority of novel algorithms while comparing them to existing state-of-the-art algorithms. It is also interesting to note that users of the website *Spliddit*, the platform from which data was used by Gal et al. [19], are provided with a detailed explanation on why the proposed rent division is fair, thus signifying the utility of explanations in everyday usage. Lee et al. [35] used visualizations to both provide an elaborate breakdown of the process as well as let the participant experiment with different values in the website *Spliddit*, to observe the changes. When the participants were shown only their outcome, they perceived the results as unfair while when they were shown the preferences and outcomes for all of the participants in the group, the participants perceived the result as fair.

From Table 1 it can be observed that Normative Characterization and Contrastive Explanations are extensively used in comparison to their other theoretical and behavioral counterparts, respectively.

6 Evaluation Methods

There are several dimensions for evaluating methods of Explainable Mechanism Design from theoretical as well as practical perspectives. We provide a description on each of them in the following paragraphs.

Theoretical Properties. Building explanations on the foundation of concepts like Axiomatic Characterization, Logic-based Programming and Automated Reasoning necessitates these methods to be supported by rigid theoretical norms. Examples include proof of an explanation given the problem instance [13,68], uniqueness of the outcome and justification given a voting profile and normative basis [9], the correctness of a tableau-based calculus for generating explanations [10] and an upper bound on the number of steps required to generate justifications to ensure readability [55].

Computational Complexity. Practical feasibility of any explanation-generation method is tantamount to its real-life application. Very few authors have addressed this aspect in their work. Exceptional examples include, Peters et al. [54] that demonstrate the polynomial-time verifiability of their proposed heuristic algorithm. Boixel and de Haan [12] prove the intractability of finding and generating justifications given a normative basis.

Empirical Analysis. Empirical insights act as an intermediate between theoretical guarantees and experimental results. Running the explainability studies on real or synthetic datasets can help compare the running times of several explanation generation methods and pick the fastest one [45], help understand the step-by-step breakdown of the explanation generation method [55], disclose interesting insights about different statistical cultures (e.g., probability distribution of election profiles) that might help in the development of personalized explanations [45], and identify and evaluate metrics for the evaluation of explanation-generation methods before deploying them in real-time studies [21].

User Studies. User studies are an effective means for examining the consequences of explanations on aspects such as reliability, satisfaction, trust and conviction. While it is always desirable to conduct the experiment with the actual participants of a mechanism as in the case of *Spliddit* [19], experiments conducted with synthetic data using platforms such as Amazon Mechanical Turk (AMT) or laboratory settings are a great starting point [48,65].

In addition to the obvious purpose of helping determine the impact of explanations, user studies can also be used to provide insights on curating effective explanations. For example, Suryanarayana et al. [65] hinted at a user bias in favor of plurality voting rule while Mosca and Such [43] used the experimental insights to devise a hybrid explanation framework and improve the wording of the explanations.

From Table 1 we can observe that there is a good mix of all of the evaluation methods in the literature. However, performing *Empirical Evaluation* and *User Studies* to evaluate explanation-generation methods in mechanism design is challenging and we address this issue in detail in the next section.

7 Challenges and Possible Solutions

Despite the rapid increase in interest in explainable systems for mechanism design, the progress made in this field is still far behind compared to XAI. One of the main challenges is that of testing. As far as testing for the efficacy of explanations is considered, the ideal premise would be testing with real users, as in *Spliddit* [19] (where explainees are the actual renters in a rent division setting) or in the work of Pozanco et al. [56] (where a real restrictive *return to office* scenario due to the COVID-19 pandemic was tackled). However, the development of such evaluations is expensive and users of such real-life settings are often inaccessible to the research community. We therefore outline a few challenges in designing

nearly realistic experiments that can be conducted in laboratories or platforms like AMT.

User Behavior. Human participants in a mechanism are poles apart from the perfect agents modeled in theory and exhibit short-sighted and downright irrational behavior. Instances of such behavior include reward divisions that adhere to weaker axioms than those that characterize Shapley value [16], playing dominated strategies in cooperative settings such as fair division and bargaining [30], and performing manipulations that can be captured by simple heuristics [40]. Any *social* explanation [41] thus catering to the expected selfish interests, while the participants betray the same, may defeat the purpose of explanations. For example, in an experiment conducted on human behavior in voting, Tal et al. [66] report that the participants exhibit herding by disregarding their most preferred candidate and voting for the candidate with the most first place votes in a predictive poll, even though it is the least optimal choice for them. In this case, framing a contrastive explanation comparing the participant's most preferred candidate and the winning candidate (which might be the candidate the participant voted for) would be counter-productive.

A natural solution to the problem of mismatched behavior is building predictive models using behavioral, game-theoretic and machine learning tools. Examples include models for predicting human decisions in plurality voting [17], approval voting [61] and auctions [50]. The benefits of such models are twofold. First, it might help the explanations to be more *selective* [41] by shedding light on what is important to the explainee. For example, in the game-theoretic model of human behavior in Doodle Polls [52], the concept of *Social Bonus* is proposed in order to reason why voters vote for unpopular slots. Consequently, contrastive explanations comparing the winning candidates to the unpopular ones can be discarded as the latter are insignificant to the voter.

The second and rather consequential utility from such models is that they might help identify the sub-optimal manipulations of the participants which can be contrasted with the optimal choice. In that context, an interesting hypothesis to investigate is if and what kind of explanations can bring irrational humans closer to the rational agents modeled in literature. This will open new avenues for *Interpretability* in mechanism design which has not received as much attention as in the XAI literature [59].

User Biases. The prevelance of mechanisms in society has led to humans forming their own prejudices such as favoritism for plurality voting rule [38,55], altruism towards non-performing participants [16] and a disdain towards algorithmic decisions as being far from reality [34,39,67].

Long before building explainable systems was considered, researchers invested efforts into manually explaining technical jargon to non-expert participants. Notable examples include acquainting participants of a centipede game with backward induction [39] and measuring the frequency of violation of fairness criteria in voting [38]. Coupling these ideas with biases such as *automation bias*, where a user believes that a computing system is more knowledgeable and

"intelligent" than it is, is a direction worth exploring [22]. In addition, comparing different modes of presenting explanations such as textual and visual, both of which are extensively used in Explainable Recommender Systems [69], can strengthen the efficacy of explanations. It is also noteworthy that human intelligence can be leveraged to not only rate explanations but also to provide explanations, thus providing valuable insights into human factors that might be useful for generating convincing explanations [56,65] .

Lack of Data. A useful tool in bridging the gap between idealized agent behavior and flawed human behavior is empirical analysis and subsequent modeling of human participants in the different mechanism design settings. Also, as mentioned earlier, empirical analysis can act as an intermediate stage between theoretical guarantees and user studies while revealing interesting insights.

However, there are not many datasets in the domains of Computational Social Choice and Preference Reasoning publicly available [37]. Also, while collecting, preserving and presenting data on private preferences, adequate care needs to be taken to ensure that user privacy is preserved [29].

Naturally, the obvious solution to the lack of data is to develop tools for efficient data collection. One such very useful collection of datasets in the domain of Computational Social Choice is *Preflib*[4] which was used by Nardi [45] to examine the practical utility of the proposed algorithm. However the process of data collection is easier said than done. Replicating real-life settings in order to get people to report their preferences, even manipulated ones, is a mammoth task. Anonymizing data is an effective way to protect the privacy of the participants. An alternate technique to preserve privacy was used by Gal et al. [19] where the original valuations for the rooms were perturbed by an acceptable margin and presented to the explainees.

Simulating Synthetic Environments. In order to evaluate the efficacy of explanations, it is vital to have the participant interested in the explanations. In mechanism design, these interests are captured by notions such as preferences, utility and costs which are easy to conceptualize but difficult to replicate and/or induce in lab experiments. XAI, even though tasked with explaining complex algorithms, enjoys relatability with experiments such as image classification [51], review classification [31] and selection of a competent agent [3]. This enables the design of interactive experiments where explanations can be sneaked in without being explicit, hence eliciting an organic response from the participant.

Inspired by the experimental design in XAI, gamification of the problems is a good starting point. Tailoring games such as the centipede game for bargaining [39] and share-the-loot game for resource allocation [16] to accommodate explanations and with the reward tied to the performance of the participant is an idea worth testing. The presence of a monetary reward inadvertently engrosses the participant, thus eliciting a realistic response. Some other tested methods of invoking user interest in synthetic lab experiments were done using the concept

[4] https://www.preflib.org/

of bonus from the winning candidate in ranked choice voting [65] and asking the participant to imagine themselves in the setting and extracting their preferences through meticulously designed questionnaires [43].

Another way to stimulate the interest of participants in explanations might be to leverage the diversity of axioms to build argumentation systems augmented with human input on how convincing the arguments are [58]. The conversion from passive listeners of explanations to active debaters of arguments might trigger a passionate yet honest response from the participants.

In addition to the above ideas, tools used in Social Psychology such as *Experimental Vignette Methodology (EVM)* [1] and online testing methods like A/B testing used in Explainable Recommendation [69] offer valid premises for developing tests for Explainable Mechanism Design.

8 Conclusion

In this paper, we survey explainability in mechanism design, provide an overall picture of the various concepts around it and shed light on the challenges faced by researchers in the domain.

While we do propose several workarounds to overcome the aforementioned challenges, we emphasize that implementing each of these is a non-trivial task per se and calls for collaborations between researchers in mechanism design, human-agent interaction, software engineering, and psychology. We hope that both experienced as well as budding researchers find this survey helpful in designing and improving explainability in mechanism design. We also envision a future where designing mechanisms aligned with human values and Explainable Mechanism Design complement each other.

Acknowledgements

This work was supported in part by the Data Science Institute at Bar-Ilan University, the EU Project TAILOR under grant 952215 and the Israeli Ministry of Science & Technology under grant 89583. The research was carried out with the technological support and funding from the HRI Consortium – the Israel Innovation Authority. Sharadhi Alape Suryanarayana is grateful for the President's Scholarship and Erasmus+ Global Mobility Programme that has supported this research.

References

1. Aguinis, H., Bradley, K.J.: Best practice recommendations for designing and implementing experimental vignette methodology studies. Organizational research methods **17**(4), 351–371 (2014)
2. Ahani, N., Gölz, P., Procaccia, A.D., Teytelboym, A., Trapp, A.C.: Dynamic placement in refugee resettlement. arXiv preprint arXiv:2105.14388 (2021)

3. Amir, O., Doshi-Velez, F., Sarne, D.: Summarizing agent strategies. JAAMAS **33**(5), 628–644 (2019)
4. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion **58**, 82–115 (2020)
5. Arrow, K.J.: Social choice and individual values. Yale university press (1951)
6. Barwise, J., Etchemendy, J.: Visual information and valid reasoning. In: Logical reasoning with diagrams (1991)
7. Belahcene, K., Chevaleyre, Y., Maudet, N., Labreuche, C., Mousseau, V., Ouerdane, W.: Accountable approval sorting. In: IJCAI-ECAI. pp. 70—76 (2018)
8. Benabbou, N., Chakraborty, M., Ho, X.V., Sliwinski, J., Zick, Y.: Diversity constraints in public housing allocation. In: AAMAS. pp. 973—981 (2018)
9. Boixel, A., Endriss, U.: Automated justification of collective decisions via constraint solving. In: AAMAS. pp. 168–176 (2020)
10. Boixel, A., Endriss, U., de Haan, R.: A calculus for computing structured justifications for election outcomes. In: AAAI (2022)
11. Boixel, A., Endriss, U., Nardi, O.: Displaying justifications for collective decisions. In: IJCAI (July 2022), demo Paper
12. Boixel, A., de Haan, R.: On the complexity of finding justifications for collective decisions. In: AAAI. pp. 5194–5201 (2021)
13. Cailloux, O., Endriss, U.: Arguing about voting rules. In: AAMAS. pp. 287–295 (2016)
14. Carroll, G.: Design for weakly structured environments. In: The future of economic design, pp. 27–33. Springer (2019)
15. Chakraborti, T., Sreedharan, S., Kambhampati, S.: The emerging landscape of explainable automated planning & decision making. In: IJCAI. pp. 4803–4811 (2020)
16. d'Eon, G., Larson, K.: Testing axioms against human reward divisions in cooperative games. In: AAMAS. pp. 312–320 (2020)
17. Fairstein, R., Lauz, A., Meir, R., Gal, K.: Modeling people's voting behavior with poll information. In: AAMAS. pp. 1422–1430 (2019)
18. Freedman, R., Borg, J.S., Sinnott-Armstrong, W., Dickerson, J.P., Conitzer, V.: Adapting a kidney exchange algorithm to align with human values. In: AAAI. pp. 1636—1645 (2018)
19. Gal, Y., Mash, M., Procaccia, A.D., Zick, Y.: Which is the fairest (rent division) of them all? In: ACM EC. pp. 67–84 (2016)
20. Garfinkel, S., Matthews, J., Shapiro, S.S., Smith, J.M.: Toward algorithmic transparency and accountability. Communications of the ACM **60**(9), 5–5 (2017)
21. Georgara, A., Rodriguez-Aguilar, J.A., Sierra, C.: Building contrastive explanations for multi-agent team formation. In: AAMAS (2022)
22. Goddard, K., Roudsari, A., Wyatt, J.C.: Automation bias: a systematic review of frequency, effect mediators, and mitigators. JAMIA (2012)
23. Goldman, J., Procaccia, A.D.: Spliddit: Unleashing fair division algorithms. ACM SIGecom Exchanges **13**(2) (2015)
24. Hart, S.: Shapley value. In: Game Theory, pp. 210–216. Springer (1989)
25. Hurwicz, L.: Optimality and informational efficiency in resource allocation processes. Mathematical methods in the social sciences (1960)
26. Kirsten, M., Cailloux, O.: Towards automatic argumentation about voting rules. In: APIA (2018)
27. Knapp, D.L.: Justification of matching outcomes. Ph.D. thesis, Master's thesis, ILLC, University of Amsterdam (2022)

28. Kominers, S.D.: Good markets (really do) make good neighbors. ACM SIGecom Exchanges **16**(2), 12–26 (2019)
29. Kraus, S., Azaria, A., Fiosina, J., Greve, M., Hazon, N., Kolbe, L., Lembcke, T.B., Muller, J.P., Schleibaum, S., Vollrath, M.: Ai for explaining decisions in multi-agent environments. In: AAAI. pp. 13534–13538 (2020)
30. Kyropoulou, M., Ortega, J., Segal-Halevi, E.: Fair cake-cutting in practice. In: ACM EC. pp. 547–548 (2019). https://doi.org/10.1145/3328526.3329592, https://doi.org/10.1145/3328526.3329592
31. Lai, V., Tan, C.: On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In: ACM FAccT. pp. 29–38 (2019)
32. Langley, P.: Explainable, normative, and justified agency. In: AAAI. pp. 9775–9779 (2019)
33. Lavi, R.: Mechanism design. Complex Social and Behavioral Systems: Game Theory and Agent-Based Models pp. 317–333 (2020)
34. Lee, M.K., Baykal, S.: Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In: CSCW. pp. 1035–1048 (2017)
35. Lee, M.K., Jain, A., Cha, H.J., Ojha, S., Kusbit, D.: Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. CSCW pp. 1–26 (2019)
36. Lee, M.K., Kusbit, D., Kahng, A., Kim, J.T., Yuan, X., Chan, A., See, D., Noothigattu, R., Lee, S., Psomas, A., et al.: Webuildai: Participatory framework for algorithmic governance. Proceedings of the ACM on Human-Computer Interaction **3**(CSCW), 1–35 (2019)
37. Mattei, N.: Closing the loop: Bringing humans into empirical computational social choice and preference reasoning. In: IJCAI. pp. 5169–5173 (2021)
38. McCune, D., McCune, L.: How can we compare different voting methods? a voting theory project. PRIMUS **29**(5), 487–501 (2019)
39. McKelvey, R.D., Palfrey, T.R.: An experimental study of the centipede game. Econometrica pp. 803–836 (1992)
40. Mennle, T., Weiss, M., Philipp, B., Seuken, S.: The power of local manipulation strategies in assignment mechanisms. In: IJCAI. pp. 82—89 (2015)
41. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence **267**, 1–38 (2019)
42. Mohseni, S., Zarei, N., Ragan, E.D.: A multidisciplinary survey and framework for design and evaluation of explainable ai systems. ACM TiiS **11**(3-4), 1–45 (2021)
43. Mosca, F., Such, J.: An explainable assistant for multiuser privacy. Autonomous Agents and Multi-Agent Systems **36**(1), 1–45 (2022)
44. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: ACM FAccT. pp. 607–617 (2020)
45. Nardi, O.: A Graph-Based Algorithm for the Automated Justification of Collective Decisions. Master's thesis, ILLC, University of Amsterdam (2021)
46. Nardi, O., Boixel, A., Endriss, U.: A graph-based algorithm for the automated justification of collective decisions. In: AAMAS (2022)
47. Nisan, N., Ronen, A.: Algorithmic mechanism design. Games and Economic behavior **35**(1-2), 166–196 (2001)
48. Nizri, M., Hazon, N., Azaria, A.: Explainable shapley-based allocation. In: AAAI (2022)

49. Noothigattu, R., Gaikwad, S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P., Procaccia, A.: A voting-based system for ethical decision making. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
50. Noti, G., Syrgkanis, V.: Bid prediction in repeated auctions with learning. In: The Web Conference. pp. 3953–3964 (2021)
51. Nourani, M., Kabir, S., Mohseni, S., Ragan, E.D.: The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In: AAAI. pp. 97–105 (2019)
52. Obraztsova, S., Polukarov, M., Rabinovich, Z., Elkind, E.: Doodle poll games. In: AAMAS. pp. 876–884 (2017)
53. Papadimitriou, C.: Algorithms, games, and the internet. In: STOC. pp. 749–753 (2001)
54. Peters, D., Pierczyski, G., Shah, N., Skowron, P.: Market-based explanations of collective decisions. In: AAAI. pp. 5656–5663 (2021)
55. Peters, D., Procaccia, A.D., Psomas, A., Zhou, Z.: Explainable voting. NeurIPS **33**, 1525–1534 (2020)
56. Pozanco, A., Mosca, F., Zehtabi, P., Magazzeni, D., Kraus, S.: Explaining preference-driven schedules: the expres framework. In: ICAPS (to appear) (2022)
57. Procaccia, A.D.: Axioms should explain solutions. In: The Future of Economic Design, pp. 195–199. Springer (2019)
58. Rosenfeld, A., Kraus, S.: Providing arguments in discussions on the basis of the prediction of human argumentative behavior. ACM TiiS **6**(4), 1–33 (2016)
59. Rosenfeld, A., Richardson, A.: Explainability in human–agent systems. JAAMAS **33**(6), 673–705 (2019)
60. Roth, A.E.: The economics of matching: Stability and incentives. Mathematics of operations research **7**(4), 617–628 (1982)
61. Scheuerman, J., Harman, J., Mattei, N., Venable, K.B.: Modeling voters in multi-winner approval voting. In: AAAI. vol. 35, pp. 5709–5716 (2021)
62. Schumann, C., Foster, J., Mattei, N., Dickerson, J.: We need fairness and explainability in algorithmic hiring. In: AAMAS. pp. 1716—1720 (2020)
63. Schwartz, S.H.: An overview of the schwartz theory of basic values. Online readings in Psychology and Culture **2**(1), 2307–0919 (2012)
64. Shapley, L.S.: A value for n-person games. Contributions to the Theory of Games **28**(2), 307–317 (1953)
65. Suryanarayana, S.A., Sarne, D., Kraus, S.: Justifying social-choice mechanism outcome for improving participant satisfaction. In: AAMAS (2022)
66. Tal, M., Meir, R., Gal, Y.: A study of human behavior in online voting. In: AAMAS. pp. 665–673 (2015)
67. Uhde, A., Schlicker, N., Wallach, D.P., Hassenzahl, M.: Fairness and decision-making in collaborative shift scheduling systems. In: CHI. pp. 1–13 (2020)
68. Zahedi, Z., Sengupta, S., Kambhampati, S.: Why didn't you allocate this task to them?'negotiation-aware task allocation and contrastive explanation generation. arXiv preprint arXiv:2002.01640 (2020)
69. Zhang, Y., Chen, X., et al.: Explainable recommendation: A survey and new perspectives. Foundations and Trends® in Information Retrieval **14**(1), 1–101 (2020)