The Fairness Fair: Bringing Human Perception into Collective Decision-Making

Hadi Hosseini

Pennsylvania State University
University Park, PA, USA
hadi@psu.edu

Abstract

Fairness is one of the most desirable societal principles in collective decision-making. It has been extensively studied in the past decades for its axiomatic properties and has received substantial attention from the multiagent systems community in recent years for its theoretical and computational aspects in algorithmic decision-making. However, these studies are often not sufficiently rich to capture the intricacies of human perception of fairness in the ambivalent nature of the real-world problems. We argue that not only fair solutions should be deemed desirable by social planners (designers), but they should be governed by human and societal cognition, consider perceived outcomes based on human judgement, and be verifiable. We discuss how achieving this goal requires a broad transdisciplinary approach ranging from computing and AI to behavioral economics and human-AI interaction. In doing so, we identify shortcomings and long-term challenges of the current literature of fair division, describe recent efforts in addressing them, and more importantly, highlight a series of open research directions.

1 Introduction

Fairness is a fundamental societal principle for promoting ethical solutions in algorithmic decision-making. It has become the center of attention in the age of Internet economics (Moulin 2019, 2004) that employs collective decision-making to deal with multiple stakeholders. Arguably, one of the most notable advancements in multiagent systems is setting policies, norms, and axioms of interactions. Recent progress in this arena has called for ethical decision making (Murukannaiah et al. 2020) in multiagent applications including algorithmic hiring (Schumann et al. 2020), blockchain technology (Grossi 2022), and in general computational social choice (Lang and Rothe 2016; Amanatidis et al. 2022a).

Within computational social choice, *fair division* has emerged to provide practical solutions for distributing goods, services, or tasks in a wide array of problems that people face in their daily lives. Its primary objective is promoting social values such as fairness and welfare with provable axiomatic guarantees. Fair division has provided a rich framework for studying mathematical foundations of societal good and has enabled advancements in algorithmic solutions that adhere to fairness concepts such as envy-freeness. A number of fair algorithms, particularly in allocation of *indivisible* items, have been made available to the public for real-world decisions through efforts such as Spliddit (www.spliddit.org), MatchU.ai (www.matchu.ai), and Course Match (University of Pennsylvania).

Over the past decades, a plethora of approximate fairness axioms have been proposed to escape negative results driven by i) impossibilities of guaranteeing certain fairness notions, ii) computational intractability, and iii) incompatibility with other socially desirable notions. Despite the rigorous research in this area, to date, the choice of which fairness axiom (or approximations) to choose has remained largely unexplored, leaving social planners uncertain as to which algorithmic solutions to adopt in practice. Furthermore, the advent of social platforms, online market places, and participatory resource allocation (Lee et al. 2019b) has shifted its attention from focusing on agents powered by artificial intelligence (AI) to human-human or human-AI interactions (Endriss 2017; Brandt et al. 2016). Thus, recent efforts in collective decision-making (d'Eon and Larson 2020; de Clippel and Rozen 2022; Nizri, Azaria, and Hazon 2022) have focused on empirical validation of axioms in computational social choice. The standard theoretical and algorithmic models of fair division often do not capture the intricacies of real-world settings that are shaped by individuals' cognitive abilities, the availability of information, and more importantly, the community and human psyche. Hence, a fundamental question emerges as to which axioms of fairness (or relaxation thereof) are perceived to be fairer? And why?

Overview. In this paper, we argue for a broader agenda in fair division based on perceived fairness, one that grounds fairness judgements based on human values.[1] With the popularity of axiomatic approach in online systems (Tennenholtz, Zohar, and Moulin 2016), Procaccia (2019) articulates—through two examples rooted in fair division and voting—that axioms should explain solutions beyond providing

---

[1]These new directions may be applied to a broader set of problems in computational social choice (e.g. in voting or cake-cutting). Here, we primarily focus on fair allocation of indivisible resources for the ease of exposition.
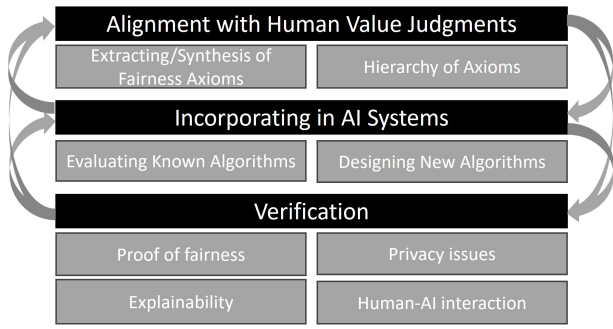
Figure 1: A high-level steps in the fairness axiom lifecycle.

mathematical guarantees. We further postulate that not only axioms of fairness should explain solutions, but they should be 1) aligned with societal and cognitive values, 2) incorporated in AI systems that are acceptable by stakeholders (e.g. individuals, communities) and social planners (and policy makers), and 3) be verifiable, i.e. individuals (or collectives) should be able to verify the solutions (see Figure 1).

Bringing human perception into decision-making leads to furthering our understanding of societal fairness when deploying algorithms based on various fair solutions, and ultimately leads to the development of new axioms of fairness. Understanding what is fair and the perception of fairness for individuals or a society of agents requires a transdisciplinary approach from AI and computing research to experimental economics and behavioral psychology. Focusing on a broad perspective, we discuss challenges in incorporating human perception along the dimensions of *value judgements*, *integration*, and *verification*. Thus, the overarching questions are:

> *What axioms are more aligned with human value judgments among all known fairness axioms? What (and how) cognitive and behavioral factors influence individuals' perception of fairness, and how should these human judgements form new fairness concepts and inform the design of new algorithms?*

2 The Fairness Fair

A standard fair division problem aims at distributing a set of m indivisible resources (or goods) among a set of n agents such that the resulting solution satisfies some axioms of fairness. The preference of an agent is generally specified with a valuation function v_i that maps each 'solution' to a value. This model allows for a wide variety of preferences over the outcomes beyond the standard assumptions in fair division that often consider valuation functions to be (i) *idiosyncratic* with *no externalities* and (ii) *additive* (the utility of each individual is the sum of values of individual items).

The primary fairness notions can be seen as either *comparison-based* notions that rely on pairwise comparisons (e.g. envy-freeness and its relaxations), and *threshold-*

based criteria that deal with achieving a fair-share of the set of items (e.g. proportionality and maximin share).[2]

One of the most desirable comparison-based axiom of fairness is *envy-freeness* (EF), which requires that no agent prefers the allocation of another agent to its own (Foley 1967; Gamow and Stern 1958). A well-studied threshold-based axiom is *proportionality* that requires giving each agent to receive at least $1/n$ of its value for all the items. Its extension, maximin share (MMS), is the value an agent can guarantee by partitioning the items into n bundles and receiving the least preferred one (Budish 2011). It is crucial to note that fairness is often studied in conjunction with economic efficiency notions such as Pareto optimality to guarantee some level of social welfare.

Relaxations. When dealing with indivisible items, none of the above fairness notions can be guaranteed. For example, consider a single valuable item and two agents. Moreover, given an instance, deciding whether such an allocation exists is generally proven to be computationally intractable. These negative results have motivated deterministic relaxations of such solution concepts (in Section 4 we discuss an alternative method that utilizes randomization).

The most prominent relaxations of fairness axioms in the literature are envy-freeness up to one item (EF1) (Lipton et al. 2004), envy-freeness up to any item (EFX) (Caragiannis et al. 2019), and (approximate) maximin share fairness (MMS) (Budish 2011). While EF1 allocations always exist and can be computed in polynomial time (Lipton et al. 2004; Caragiannis et al. 2019), the existence and computation of EFX allocations—-with some exceptions for instance when $n < 4$ or under lexicographic extensions (Chaudhury, Garg, and Mehlhorn 2020; Hosseini et al. 2021)—are intriguing open problems that have motivated a large array of works in recent years. Similarly, the non-existence of MMS allocations (Kurokawa, Procaccia, and Wang 2018) has inspired a variety of algorithmic techniques in finding multiplicative (Ghodsi et al. 2021; Garg and Taki 2021) or ordinal (Hosseini, Searns, and Segal-Halevi 2022b,a) approximations to the MMS threshold.[3]

The following example illustrates an instance of a fair division problem. We will use it as a running example when discussing the challenges and directions in incorporating human perception in fair decision-making.

Example 1. *Consider three agents with preferences as shown in Table 1. For simplicity of exposition, the preferences are assumed to be additive. The allocation shown with circles is EF1, but does not satisfy proportionality: Agent 1 is envy-free; the envy of agents 2 and 3 towards agent 1 can be eliminated by the hypothetical removal of a single item (g_3). The allocation shown with underlines is likewise EF1*

---

[2]We do not discuss notions that measure the degree of fairness in a society by minimizing the maximum or sum of envy (Chevaleyre et al. 2007; Chen and Shah 2018; Nguyen and Rothe 2014), minimizing the envy ratio (Lipton et al. 2004), or balancing the amount of envy experienced in a society.

[3]For a comprehensive list, see the following recent surveys (Aziz et al. 2022; Amanatidis et al. 2022b).

	g_1	g_2	g_3	g_4
1	⑥	4	⑩	10
2	7	6	9	⑧
3	1	⑪	12	6

Table 1: A stylized running example with four items and three agents.

but improves the economic efficiency (more precisely, it is EF1 and Pareto optimal).

3 Alignment with Human Value Judgement

In the field of fair division, the most common fairness axioms were originated from the early works of mathematicians such as Steinhaus on fairly dividing a cake (Steinhaus 1948) and insights that followed from philosophical theories in distributive justice (Rawls 1971; Sen 2018). Over the next few decades, these studies resulted in the development of a plethora of fairness axioms. While these mathematical constructs gave rise to advancements in mechanism design and algorithmic solutions, they do not completely capture the intricacies of human decision judgement, as often fairness axioms (or outcomes) themselves influence the preferences of individuals (Sandbu 2008).

In what follows, we argue that 1) the well-studied fairness concepts should be revisited and evaluated in the presence of complexities in human perception, 2) it is critical to investigate competing fairness axioms and their (in)compatibility when interacting with individuals, and 3) there is a dire need for devising a *hierarchy of axioms* in fair decision-making to ensure that the axioms are aligned with human preferences.

3.1 Extracting Fairness and Hierarchy of Axioms

A compelling argument for the aforementioned fairness concepts is that they do not rely on *interpersonal* utility comparisons[4], meaning that they eliminate the need for identifying who derives the most (or the least in case of chores/tasks) from (a bundle of) resources. Thus, an agent values its own bundle *only* according to its own direct *introspective* preferences. Yet, in reality human value judgement is seldom introspective and is often affected by a variety of individual, social, and cognitive parameters. This raises the questions of *which fairness axioms are perceived to be fairer according to human values? Do envy-free allocations (when they exist) are perceived to be fairer compared to their proportional counterparts? What about their relaxations? Which fairness axioms are more aligned with human value judgements?*

3.2 Perceived Fairness and Cognitive Aspects

In this section, we highlight a few (non-exhaustive) factors that are often involved in human judgement and discuss their impacts on algorithmic fairness.

---

[4]For a comprehensive discussion on interpersonal utility comparisons and arguments for and against them in the utility theory, see (Sen 2018; Harsanyi 1990).

Transparency and Information Sharing. Counterfactual reasoning seems to be the cornerstone of several approximate fairness concepts. Many prominent relaxations of fairness such as EF1 and EFX rely on a counterfactual interpretation of fairness, assuming that an agent's envy is eliminated if a single item is to be removed from the envied agents. While transparency is a crucial societal concern, in practice often agents do not have access to all the information. In this vein, *epistemic* relaxations of fairness have been proposed that rely either on *withholding* information (Aziz et al. 2018) or on removal of a randomly selected item (Farhadi et al. 2021). A question is *how to utilize information sharing/withholding to increase the perceived fairness of outcomes?*

A recent epistemic relaxation is *envy-freeness up to k hidden goods* (HEF-k) that eliminates envy by hiding a small (k) number of items (Hosseini et al. 2020). Interestingly, an experimental study showed that HEF-k allocations are perceived to be fairer compared to their EF1 counterparts (Hosseini et al. 2022). Yet, these studies are limited in scope as they pertain to specific settings with highly stylized instances. Moreover, it is not clear how information sharing guide (or cause) an outcome to be perceived fairer. One plausible justification could revolve around individuals' bounded rationality or the amount of effort needed to consider the 'unknown', which has been described by Daniel Kahneman as '*what you see is all there is*'—the cognitive phenomenon that human brain is hardwired to believe that the given information is the entirety of the relevant information.

Fair for Me vs. Fair for All. In fair division, agents make judgements over their own allocations, either in comparison with what is allocated to other agents (as in envy-freeness) or by measuring its value against a fairness threshold (as in proportionality). However, in reality fairness judgments are often made based on human values for fair outcomes *for all*, meaning that what is fair for me may not be considered as a fair outcome for everyone.

In the context of *bargaining*, Herreiner and Puppe (2009) demonstrated experimentally that the primary factor of fairness is in fact *inequality aversion*—i.e. the tendency of humans in seeking equal outcomes—and that envy-freeness plays a secondary role in the perception of fairness, only when economic efficiency and inequality aversion are satisfied. In other words, everything equal, people tend to associate a higher order to inequality aversion. These findings suggest that individuals seem to (implicitly) create a hierarchy of axioms to reason about relative value of each outcome. Hence, developing fair algorithms may require axioms that are able to capture solutions that take the social context into account beyond perceived envy.

Let us illustrate this point by visiting Example 1: Agent 1 may not perceive any unfairness in the circled allocation as she does not envy others, however, she may perceive the overall outcome to be unfair since she receives a higher value (as well as higher number of items) than the other two agents. Does agent 1 consider the outcome fair for agent 3 since she receives her proportional share? More broadly, *which fairness concepts are employed in assess-*

ing outcomes? And do those fairness concepts change when evaluating an outcome for self vs. for everyone?

Skin in the Game. When measuring fairness of an outcome for everyone, there is an element of how much stake an agent has in the decision itself. In other words, human judgement is often dependent on whether (and to what extent) they have 'skin in the game'. Looking into the fairness notions such as envy-freeness, it is crucial to understand the perception of fairness both from the eyes of participating agents and non-participating *bystanders*. This concept is related to avoiding 'self-serving bias' in behavioral economics (Konow 2003). Here, the goal is to understand the extent of which having 'skin in the game' can impact human perception, and whether fairness axioms should be sensitive to such biases. *Would an envy-free agent consider the outcome fair if they were to observe it as a bystander?*

One approach is to use the judgement of the other individuals as a proxy for objectivity by measuring fairness in 'the eyes of the others' (Shams et al. 2022); thus, envy exists if sufficiently many agents agree that it should be the case (if they were in his shoes). It is a generalization of the unanimous envy condition proposed by Van Parijs (1995). An open question is, thus, whether individuals tend to value the views of others (and perhaps those closer to them) when assessing the fairness of outcomes.

Agency & Deliberation. Given the key role of 'agency' in human perception, an intriguing questions arises on whether (and to what extent) distributed fair algorithms are perceived to be fairer compared to centralized algorithms (see the following discussion on procedural vs. distributive fairness). The impact of agency through outcome control, i.e. the ability to adjust the outcomes, has shown to improve the perceived fairness among participants in resource allocation studies (Lee et al. 2019a). More importantly, even when allocations satisfy the fairness notion of EF1, the sheer 'sense of control and agency' resulted in higher satisfaction even when modifications were performed as a group. In voting, Grandi et al. (2022) has recently shown that group deliberation in a repeated setting improves the social outcome where participants demonstrated a more optimistic behavior when given agency to deliberate. Similar experiments in fair resource allocation illustrate that agency and group deliberation improve participants perception of fairness in comparison with outcomes prescribed by algorithms on Spliddit (Lee and Baykal 2017). In the context of matching, a recent experimental study, similarly, illustrates the role of autonomy and agency in perceived fairness of matching algorithms (König et al. 2019).

Thus, the following questions arise: *How (and to what extent) involving participants in the decision-making process improves their perception of fairness? And how the perceived fairness relates to human cognitive biases such as the Ikea Effect (Norton, Mochon, and Ariely 2012)— tendency to associate higher value to products/outcomes one contribute—and cognitive dissonance, as observed by Konow (2003)?*

Motivation, Context, and Framing. Individuals' motivation for a task as well as the context (e.g. culture) impacts their actions and perceptions. A recent study on resource allocation (Lee 2018) suggests that the perceived fairness of algorithms is *dependent on task characteristics*; and participants consider algorithmic solutions to be equally fair in mechanical tasks (e.g. processing data) compared to solutions suggested by humans, while they perceive algorithm-induced outcomes to be less fair for tasks that require human skills and subjective judgement.

Moreover, the sensitivity of fairness axioms to the well-known *framing effect* (Tversky and Kahneman 1985)– whether in evaluating outcomes or during the elicitation of preferences–could impact the choice of fairness notions in collective decision-making. Framing effect often causes *preference reversal* which states that a change in reference point may lead to a change in preference (aka preference reversal) even if the values associated with possible outcomes remain unchanged. Kahneman, Slovic, and Tversky (1982) showed that human behavior rarely adheres to a closely-related axiom, often known as the independence axiom (aka Independence of Irrelevant Alternative).

Thus, an immediate direction is studying how context and framing could impact the perception of fairness. The recent focus in studying allocation algorithms pertaining to goods (positively valued items) versus chores (negatively valued items) and mixture thereof, albeit necessary, is primarily theoretical. For example, axioms such as EF1 and EFX can easily be interpreted when distributing tasks by considering the removal of item from one's own bundle; yet, the perceived fairness of these solutions may be impacted by human's behavioral change when facing gains vs. losses—a seminal concept in *prospect theory* (Kahneman and Tversky 1979).

Distributive vs. Procedural Justice. The vast majority of research in fair division can be situated within *distributive justice*[5], which is primarily concerned with socially fair 'outcomes'. However, the human judgement of decisions is often impacted by perceived fairness of the procedures. In other words, the interaction between distributive fairness and *procedural justice* (Tyler and Allan Lind 2002) in allocation of resources—particularly when dealing with approximate fairness axioms—requires an in-depth investigation.

Let us revisit the instance given in Example 1: both circled and underlined allocations satisfy EF1; however, the circled allocation may be perceived to be fairer as it is the outcome of a fair procedure (in this case the Round-Robin algorithm): agents pick their favorite items among the remaining items one by one according to the $1, 2, 3, 1$ ordering. More importantly, note that the underlined allocation is theoretically more desirable as it results in higher social welfare (it is simultaneously EF1 and Pareto optimal).

In the context of allocating a divisible resource, procedures that are envy-free are perceived to be fairer by participants than those that are proportional (but necessarily not envy-free) (Kyropoulou, Ortega, and Segal-Halevi 2022).

---

[5]Distributive justice has a long history in philosophy within Rawls' theory of justice in distribution of social goods. We refer the readers to the seminal works of Adams (1963) and Rawls (1971).

These findings imply that participants–all things equal–can correctly identify (and prefer) procedures that provide stronger fairness guarantees. Yet, the comparison between fair procedures and fair outcomes remains an interesting open problem. *Are envy-free outcomes through centralized algorithms perceived to be fairer compared to fair procedures that only guarantee weaker notions of fairness (e.g. proportionality)?*

4 Uncertainty and Temporal Effects

Thus far we mainly highlighted challenges in devising and evaluating axioms of fairness in isolated, static, settings. Yet, human judgement is often impacted by uncertainty of the outcomes or procedures as well as dynamic or temporal nature of parameters or decisions.

Fair Lotteries. Randomization has emerged as an alternative approach to approximate fairness notions such as EF1. The goal is often to achieve fairness ex-ante by allowing agents to participate in fair lotteries. While recent efforts investigated theoretical boundaries of achieving ex-ante envy-freeness and ex post EF1 solutions (Freeman, Shah, and Vaish 2020; Aziz 2020), the interaction between the two approach and how they impact the perception of fairness remain mainly open.

Sequential Fairness. When items need to be allocated sequentially, for example in charity donations, human perception may be affected by its lookahead reasoning, different discounting factors for future allocations, as well as availability bias. Similarly, in problems involving repeated allocation (or reallocation) of resources, the following questions may arise: Which axioms of fairness can adequately capture human judgements? And how should these solutions be implemented to improve perceived fairness?

Dynamic Settings. Imagine a scenario where new individuals arrive after an allocation decision is made. What modifications or redistribution of resources are considered socially acceptable (and fair)? Similarly, when an agent leaves its items behind, how shall we redistribute its share among the rest of agents so as to maintain fairness? What constitutes a fair redistribution? Would the initial allocation's (un)fairness impact how the resdistribution should happen? For example, consider an envy-free allocation. An agent leaves and the resources it held now needs to be redistributed. Suppose redistribution A preserves envy-freeness, but another redistribution, say B, does not. Which redistribution is perceived to be fairer? What if the initial allocation is not envy-free?

5 Verification of Fairness

5.1 Proof of Fairness

A key criterion in adopting algorithmic approaches for solving societal problems is human perception. The majority of algorithms and procedures were not traditionally designed with human perception in mind. In particular, fairness as distributive justice (aka, outcome fairness) is challenging to verify outside of academic and AI expert circles; we cannot expect from people the level of algorithmic literacy that is necessary for understanding and verifying solutions. Even if we do, it is unreasonable to presume sufficient cognitive (or computational) dedication to verify the outcomes generated by algorithms. Note that even when dealing with AI-powered agents working on behalf of firms, institutions, or individuals, solutions (and their certificates) are ultimately interpreted by human stakeholders. Thus, a pressing research agenda is how to provide a *proof of fairness* that takes human perception into account.

On the philosophical front, proof of fairness is closely related to *interactional justice* (Bies and Moag 1986), that focuses on explanations that are provided to individuals to justify mechanisms that are used to ensure procedural fairness or delineate the reasons behind distribution of resources in a certain fashion.[6] These issues, at heart, require a thorough understanding of human perception, human-AI interaction, and algorithmic thinking.

5.2 Privacy-Preserving Verification

Providing a proof of fairness to participants may have severe privacy ramifications. For example, verifying envy-freeness requires sharing information about the bundles allocated to all other agents. Thus, when sensitive information needs to be shared to ensure fairness (e.g. the availability of an agent, location information in facility location problems, and evaluating colleagues or team members in an organization), there is a need for properly balancing privacy concerns with the transparency required to verify fairness of an outcome. A recent work on differential privacy in fair allocation of indivisible goods (Manurangsi and Suksompong 2022) show mainly negative results in achieving approximate fairness notions of envy-freeness and proportionality. An overarching question is *how should we devise verification mechanisms that allow for reasonable and practical proof of fairness while ensuring some level of privacy?*

A promising direction is potentially focusing on information withholding approaches (as we discussed in Section 3.2). However, these approaches may in turn have negative impact on perceived fairness. Another practical approach for privacy-preserving fairness verification is to exploit randomization techniques. For instance, *should we enable the individuals to randomly select (a limited number of) bundles to verify fairness (e.g. envy-freeness) instead of sharing information about other agents? How would this approach in return impact individuals' perception of fairness or more importantly their trust in algorithmic decision-making?*

5.3 Explainable Fairness

Taking an AI standpoint, there has been much interest in explainable AI within the machine learning community which primarily focuses on transparency in the mechanics of the machine learning algorithms (Burkart and Huber 2021;

---

[6]Interactional justice has two components: *interpersonal justice*, which focuses on treating individuals with dignity and respect by social planners or authorities, and *informational justice*, which is focused on the explanation of outcomes or procedures (Greenberg 1990, 1993). Here, we are primarily referring to the latter.

Chakraborti, Patra, and Noble 2020). We advocate for a broader sense of explainablity, that is, solutions must be explainable not only to experts and system designers but they should also provide justifications to the participating parties.

Explainability in mechanism design and voting has attracted some attention in recent years primarily from a theoretical point of view. Explainable mechanism design often has to deal with unique challenges (in contrast to the general explainable AI framework) due to complexities such as multi-objective norms, different stakeholders, and the range of the required justifications (Suryanarayana, Sarne, and Kraus 2022). In the context of voting, recent works focus on providing justifications for a given decision through a set of norms and axioms that are 'collectively accepted' by the agents along with arguments that explain the solutions (Boixel and Endriss 2020; Nardi, Boixel, and Endriss 2022). While the size of explanations can be bounded in certain domains of voting (Peters et al. 2020), how individuals measure and evaluate these explanations remains mainly unstudied. In addition, axioms of fairness rely on a variety of social norms and cognitive biases that could mutually interact with the framing, size, or the ordering of the explanations.

5.4 Human-AI Interactions and Visualization

A primary challenge in designing fair solutions is understanding how people and AI systems interact with one another. These interactions require a reasonable (and sufficient) methods to elicit complex preferences as well as communicating the outcomes (or potential outcomes) to the individuals.

Preference Elicitation. In fair allocation of indivisible items, we often consider restricted preferences (e.g. additive valuations) to allow for compact representation of preferences over bundles. While these restrictions have origins in practical applications, it is unreasonable to expect that individuals can assign *precise* cardinal valuations to a large number of items. On the other hand, ordinal preferences may not contain sufficient information required for total extensions over outcomes.

Preference elicitation in systems involving human participants requires going beyond theoretical models that aim at minimizing the number of queries. Rather, it requires proper mechanics about how to (and to what extent) interact to prevent cognitive overload and maintain a balance between optimal elicitation and achieving socially desirable decisions.

Visualizations. Visualization of the fair algorithms such as Spliddit (Shah 2017) and MatchU.ai (Ferris and Hosseini 2020) could improve the algorithmic literacy of individuals. Such efforts have been shown to improve algorithmic thinking and providing (limited) ability for comprehending certain axioms (Bao and Hosseini 2023). Yet, these visualizations are insufficient in providing proof of fairness that can then be simply verified by individuals. Thus, the verification of distributive justice (outcome fairness) not only requires fairness axioms to be coherent with human values, but also calls for equipping individuals with smooth visual interactions to check and ultimately accept the solutions.

6 Conclusion

Fairness issues in AI systems arise from a variety of axiomatic, algorithmic, and epistemic assumptions. These challenges require a precise and in-depth interdisciplinary discussions to align fairness axioms with human value judgements, evaluate the perceived fairness of common fairness notions, and develop new axioms of fairness. Novel algorithmic approaches should enable individuals to verify the fairness of the solutions by providing explanations for the procedures, enabling interactions with the algorithms, and providing a proof of fairness while preserving privacy of the individuals. From the technical perspective, new algorithms should be designed to adhere to the new fairness axioms that are aligned with human values. Computational hardness may not necessarily be an obstacle if the new axioms are shown to not rely on full information. In this vein, theoretical research may shift focus from approximate methods to context-aware solutions.

Incorporating human perception in collective decision-making is an extremely challenging task. A long-term successful plan must draw from a broad array of disciplines including AI and computing research to experimental economics and behavioral psychology. Our aim in this paper was to paint a broad agenda for future research in collective decision-making. The hope is to enrich our understanding of human judgement and develop algorithmic solutions that are responsive (or at least aware) of human perception and ultimately establish new axioms of fairness for future research.

Acknowledgments

References

Adams, J. S. 1963. Towards an understanding of inequity. *Journal of Psychopathology and Clinical Science*, 67(5).

Amanatidis, G.; Birmpas, G.; Filos-Ratsikas, A.; and Voudouris, A. A. 2022a. Fair Division of Indivisible Goods: A Survey. In Raedt, L. D., ed., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 5385–5393. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Amanatidis, G.; Birmpas, G.; Filos-Ratsikas, A.; and Voudouris, A. A. 2022b. Fair Division of Indivisible Goods: A Survey. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, 5385–5393.

Aziz, H. 2020. Simultaneously achieving ex-ante and ex-post fairness. In *International Conference on Web and Internet Economics*, 341–355. Springer.

Aziz, H.; Bouveret, S.; Caragiannis, I.; Giagkousi, I.; and Lang, J. 2018. Knowledge, Fairness, and Social Constraints. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 4638–4645.

Aziz, H.; Li, B.; Moulin, H.; and Wu, X. 2022. Algorithmic Fair Allocation of Indivisible Items: A Survey and New Questions. In *ACM SIGecom Exchanges*, volume 20, 24–40.

Bao, Y.; and Hosseini, H. 2023. Mind the Gap: The Illusion of Skill Acquisition in Computational Thinking. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, 778–784.

Bies, R. J.; and Moag, J. 1986. Interactional justice: Communication criteria of fairness. *Research on Negotiation in Organizations*, 1: 43–55.

Boixel, A.; and Endriss, U. 2020. Automated Justification of Collective Decisions via Constraint Solving. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, 168–176.

Brandt, F.; Conitzer, V.; Endriss, U.; Lang, J.; and Procaccia, A. D. 2016. *Handbook of Computational Social Choice*. Cambridge University Press.

Budish, E. 2011. The Combinatorial Assignment Problem: Approximate Competitive Equilibrium from Equal Incomes. *Journal of Political Economy*, 119(6): 1061–1103.

Burkart, N.; and Huber, M. F. 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70: 245–317.

Caragiannis, I.; Kurokawa, D.; Moulin, H.; Procaccia, A. D.; Shah, N.; and Wang, J. 2019. The Unreasonable Fairness of Maximum Nash Welfare. *ACM Transactions on Economics and Computation*, 7(3): 12.

Chakraborti, T.; Patra, A.; and Noble, J. A. 2020. Contrastive Fairness in Machine Learning. *IEEE Letters of the Computer Society*, 3(02): 38–41.

Chaudhury, B. R.; Garg, J.; and Mehlhorn, K. 2020. EFX Exists for Three Agents. In *Proceedings of the 21st ACM Conference on Economics and Computation*, 1–19.

Chen, Y.; and Shah, N. 2018. Ignorance is often bliss: Envy with incomplete information. Mansctipt.

Chevaleyre, Y.; Endriss, U.; Estivie, S.; and Maudet, N. 2007. Reaching envy-free states in distributed negotiation settings. In *Proceedings of the 20th international joint conference on Artificial intelligence*, 1239–1244. Morgan Kaufmann Publishers Inc.

de Clippel, G.; and Rozen, K. 2022. Fairness through the Lens of Cooperative Game Theory: An Experimental Approach. *American Economic Journal: Microeconomics*, 14(3): 810–36.

d'Eon, G.; and Larson, K. 2020. Testing Axioms against Human Reward Divisions in Cooperative Games. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, 312–320.

Endriss, U. 2017. *Trends in computational social choice*. AI Access.

Farhadi, A.; Hajiaghayi, M.; Latifian, M.; Seddighin, M.; and Yami, H. 2021. Almost envy-freeness, envy-rank, and nash social welfare matchings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 5355–5362.

Ferris, J.; and Hosseini, H. 2020. Matchu: An interactive matching platform. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13606–13607.

Foley, D. K. 1967. Resource Allocation and the Public Sector. *Yale Economic Essays*, 7: 45–98.

Freeman, R.; Shah, N.; and Vaish, R. 2020. Best of both worlds: Ex-ante and ex-post fairness in resource allocation. In *Proceedings of the 21st ACM Conference on Economics and Computation*, 21–22.

Gamow, G.; and Stern, M. 1958. *Puzzle-Math*. New York, NY, USA: Viking Press.

Garg, J.; and Taki, S. 2021. An Improved Approximation Algorithm for Maximin Shares. *Artificial Intelligence*, 300: 103547.

Ghodsi, M.; Hajiaghayi, M. T.; Seddighin, M.; Seddighin, S.; and Yami, H. 2021. Fair Allocation of Indivisible Goods: Improvement. *Mathematics of Operations Research*, 46(3): 1038–1053.

Grandi, U.; Lang, J.; Ozkes, A.; and Airiau, S. 2022. Voting behavior in one-shot and iterative multiple referenda. *Social Choice and Welfare*, (3747713).

Greenberg, J. 1990. Organizational justice: Yesterday, today, and tomorrow. *Journal of Management*, 16(2): 399–432.

Greenberg, J. 1993. The social side of fairness: interpersonal and informational classes of organizational justice. *Justice in the workplace: approaching fairness in human resource management*, 79–103.

Grossi, D. 2022. Social Choice Around the Block: On the Computational Social Choice of Blockchain. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, 1788–1793.

Harsanyi, J. C. 1990. Interpersonal Utility Comparisons. In *Utility and Probability*, 128–133. Springer.

Herreiner, D. K.; and Puppe, C. D. 2009. Envy Freeness in Experimental Fair Division Problems. *Theory and Decision*, 67(1): 65–100.

Hosseini, H.; Kavner, J.; Sikdar, S.; Vaish, R.; and Xia, L. 2022. Hide, Not Seek: Perceived Fairness in Envy-Free Allocations of Indivisible Goods. *arXiv preprint arXiv:2212.04574*.

Hosseini, H.; Searns, A.; and Segal-Halevi, E. 2022a. Ordinal Maximin Share Approximation for Chores. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, 597–605.

Hosseini, H.; Searns, A.; and Segal-Halevi, E. 2022b. Ordinal maximin share approximation for goods. *Journal of Artificial Intelligence Research*, 74: 353–391.

Hosseini, H.; Sikdar, S.; Vaish, R.; Wang, J.; and Xia, L. 2020. Fair Division through Information Withholding. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, 2014–2021.

Hosseini, H.; Sikdar, S.; Vaish, R.; and Xia, L. 2021. Fair and Efficient Allocations under Lexicographic Preferences. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*. Forthcoming.

Kahneman, D.; Slovic, P.; and Tversky, A. 1982. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.

Kahneman, D.; and Tversky, A. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2): 263–292.

König, T.; Kübler, D.; Mechtenberg, L.; and Schmacker, R. 2019. Fair Procedures with Naive Agents: Who Wants the Boston Mechanism? Rationality and Competition Discussion Paper Series 222.

Konow, J. 2003. Which is the fairest one of all? A positive analysis of justice theories. *Journal of economic literature*, 41(4): 1188–1239.

Kurokawa, D.; Procaccia, A. D.; and Wang, J. 2018. Fair Enough: Guaranteeing Approximate Maximin Shares. *Journal of the ACM*, 65(2): 1–27.

Kyropoulou, M.; Ortega, J.; and Segal-Halevi, E. 2022. Fair Cake-Cutting in Practice. *Games and Economic Behavior*, 133: 28–49.

Lang, J.; and Rothe, J. 2016. Fair Division of Indivisible Goods. In *Economics and Computation*, 493–550. Springer.

Lee, M. K. 2018. Understanding Perception of Algorithmic Decisions: Fairness, Trust, and Emotion in Response to Algorithmic Management. *Big Data & Society*, 5(1): 1–16.

Lee, M. K.; and Baykal, S. 2017. Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1035–1048.

Lee, M. K.; Jain, A.; Cha, H. J.; Ojha, S.; and Kusbit, D. 2019a. Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW).

Lee, M. K.; Kusbit, D.; Kahng, A.; Kim, J. T.; Yuan, X.; Chan, A.; See, D.; Noothigattu, R.; Lee, S.; Psomas, A.; et al. 2019b. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–35.

Lipton, R. J.; Markakis, E.; Mossel, E.; and Saberi, A. 2004. On Approximately Fair Allocations of Indivisible Goods. In *Proceedings of the 5th ACM Conference on Electronic Commerce*, 125–131.

Manurangsi, P.; and Suksompong, W. 2022. Differentially Private Fair Division. *arXiv preprint arXiv:2211.12738*.

Moulin, H. 2004. *Fair division and collective welfare*. MIT press.

Moulin, H. 2019. Fair division in the internet age. *Annual Review of Economics*, 11: 407–441.

Murukannaiah, P. K.; Ajmeri, N.; Jonker, C. M.; and Singh, M. P. 2020. New Foundations of Ethical Multiagent Systems. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, 1706–1710.

Nardi, O.; Boixel, A.; and Endriss, U. 2022. A graph-based algorithm for the automated justification of collective decisions. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, 935–943.

Nguyen, T. T.; and Rothe, J. 2014. Minimizing envy and maximizing average Nash social welfare in the allocation of indivisible goods. *Discrete Applied Mathematics*, 179: 54–68.

Nizri, M.; Azaria, A.; and Hazon, N. 2022. Improving the Perception of Fairness in Shapley-Based Allocations. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Norton, M. I.; Mochon, D.; and Ariely, D. 2012. The IKEA Effect: When Labor leads to Love. *Journal of Consumer Psychology*, 22(3): 453–460.

Peters, D.; Procaccia, A. D.; Psomas, A.; and Zhou, Z. 2020. Explainable voting. *Advances in Neural Information Processing Systems*, 33: 1525–1534.

Procaccia, A. D. 2019. Axioms should explain solutions. In *The Future of Economic Design*, 195–199. Springer.

Rawls, J. 1971. *A Theory of Justice*. Harvard University Press.

Sandbu, M. E. 2008. Axiomatic foundations for fairness-motivated preferences. *Social Choice and Welfare*, 31(4): 589–619.

Schumann, C.; Foster, J. S.; Mattei, N.; and Dickerson, J. P. 2020. We Need Fairness and Explainability in Algorithmic Hiring. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, 1716–1720.

Sen, A. 2018. *Collective Choice and Social Welfare*. Harvard University Press.

Shah, N. 2017. Spliddit: two years of making the world fairer. *XRDS: Crossroads, The ACM Magazine for Students*, 24(1): 24–28.

Shams, P.; Beynier, A.; Bouveret, S.; and Maudet, N. 2022. Fair in the Eyes of Others. *Journal of Artificial Intelligence Research*, 75: 913–951.

Steinhaus, H. 1948. The problem of fair division. *Econometrica*, 16: 101–104.

Suryanarayana, S. A.; Sarne, D.; and Kraus, S. 2022. Explainability in mechanism design: recent advances and the road ahead. In *European Conference on Multi-Agent Systems*, 364–382. Springer.

Tennenholtz, M.; Zohar, A.; and Moulin, H. 2016. *The Axiomatic Approach and the Internet*, 427–452. Cambridge University Press.

Tversky, A.; and Kahneman, D. 1985. The framing of decisions and the psychology of choice. In *Behavioral decision making*, 25–41. Springer.

Tyler, T. R.; and Allan Lind, E. 2002. Procedural Justice. In *Handbook of Justice Research in Law*, 65–92.

Van Parijs, P. 1995. *Real freedom for all: What (if anything) can justify capitalism?* Clarendon Press.