sustainability

MDPI

Article

Gamified Text Testing for Sustainable Fairness

Savaş Takan [1,*,†], Duygu Ergün [2,*,†] and Gökmen Katipoğlu [3]

[1] Department of Artificial Intelligence and Data Engineering, Faculty of Engineering, Ankara University, 06830 Ankara, Turkey
[2] School of Fine Arts Design and Architecture, Atılım University, 06830 Ankara, Turkey
[3] Department of Computer Engineering, Faculty of Engineering, Kafkas University, 36100 Kars, Turkey
* Correspondence: stakan@ankara.edu.tr (S.T.); duygu.ergun@atilim.edu.tr (D.E.)
† These authors contributed equally to this work.

Abstract: AI fairness is an essential topic as regards its topical and social-societal implications. However, there are many challenges posed by automating AI fairness. Based on the challenges around automating fairness in texts, our study aims to create a new fairness testing paradigm that can gather disparate proposals on fairness on a single platform, test them, and develop the most effective method, thereby contributing to the general orientation on fairness. To ensure and sustain mass participation in solving the fairness problem, gamification elements are used to mobilize individuals' motivation. In this framework, gamification in the design allows participants to see their progress and compare it with other players. It uses extrinsic motivation elements, i.e., rewarding participants by publicizing their achievements to the masses. The validity of the design is demonstrated through the example scenario. Our design represents a platform for the development of practices on fairness and can be instrumental in making contributions to this issue sustainable. We plan to further realize a plot application of this structure designed with the gamification method in future studies.

Keywords: artificial intelligence fairness; gamification; testing; fair-test; zero-suppression decision diagram; centrality; confusion matrix; mutation test

check for updates

Citation: Takan, S.; Ergün, D.; Katipoğlu, G. Gamified Text Testing for Sustainable Fairness. *Sustainability* **2023**, *15*, 2292. https://doi.org/10.3390/su15032292

Academic Editor: Hao-Chiang Koong Lin

Received: 22 November 2022
Revised: 29 December 2022
Accepted: 20 January 2023
Published: 26 January 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Justice is a multifaceted and context-dependent concept that has been debated for centuries. Specifically, for justice alone, there are many different definitions, several of which are incompatible, as no method can satisfy them simultaneously except in minimal special cases [1]. The limitations, compatibility, applicability, assumptions, and parameters of each definition are not yet fully understood [2,3].

Today's computer technologies automate decision-making in many areas. These technologies can offer tremendous opportunities, as they are accelerating scientific discovery in personalized medicine, intelligent weather forecasting, and many other regions. They can automate tasks related to personal decisions, and help improve our daily lives through personal digital assistants and recommendations. These include hiring processes, credit score prediction, and sentencing decisions in courts, as well as software systems used in surveillance. Moving forward, such decision-making has the potential to transform society through platforms such as open government. These machines are generally beneficial, as they allow for more informed and less subjective decisions that taking into account factors other than human ones. On the other hand, in the quest to maximize efficiency the decisions made by machines can be flawed [4]. In addition to making life easier, this situation has led to new debates around justice and automatization in terms of making fair decisions. In other words, justice debates have gained a new dimension today as decision-making mechanisms have shifted from humans to machines with developing automation technologies.

Maintaining fairness in emerging automation technologies means measuring and eliminating discrimination in the model and ensuring that the applications built around

the model are reliable. The goal is to prevent the model from making significantly different predictions for subgroups where each subgroup is divided by a "sensitive characteristic" such as race, gender, or age [5]. Different approaches include algorithms, data, and holistic justice [6]. These approaches often focus on discrimination issues such as gender, religion, age, sexual orientation, and race [7,8]. Because fairness is in general a difficult concept to understand, numerous definitions and models of fairness have been developed, as well as various algorithmic approaches to fair rankings and proposals that make the landscape very convoluted [9]. To make real progress in building fair systems, it is necessary to demystify what has been done, to understand how and when each model and approach can be used, and finally, to be able to recognize the research challenges that lie ahead [10]. In this respect, there is a need for holistic tests of fairness.

Text is an important building block in holistic tests for the fairness problem [11]. This is because text is an expression of language as social communication. Text is important because it carries clues about various social interaction processes. For example, when our perception is exposed to an image, we use text to process it. On the other hand, the concept of justice is structured through legal expressions, that is, through text.

Motivated by the challenges of automating fairness in texts, we aim to create a new fairness testing paradigm that combines disparate pro-fairness AI proposals into a single platform in order to test them and develop the most effective method, thereby contributing to the general direction of fairness. However, we believe that there is a need for mass participation in the best way to test texts for AI fairness. For this reason, a game for testing AI fairness in texts was designed using the gamification concept to ensure participation in the fairness system we have developed, make it sustainable, and motivate the participating individuals for this purpose.

The concept of gamification generally refers to technological, economic, cultural, and societal developments in which reality becomes more gamified. Thus, it can lead to a greater degree of skill, motivational benefits, creativity, playfulness, engagement, and overall positive growth and happiness [12]. The structuring and development of AI fairness affect technological, economic, cultural, and societal developments. The role of civic engagement in all these developments is critical [13]. In this respect, gamification practices are an important strategy for increasing civic engagement [14]. Gamification is defined as using game elements in non-game contexts to make a system more attractive by leveraging the intrinsically rewarding properties of games [15]. Significant studies show that gamified applications provide a better user experience and civic engagement [16,17].

It is known that the main reason for using gamification is to increase motivation [17]. In this way, gamification aims to change user behavior [15]. Two types of motivation for behavior change are intrinsic and extrinsic [18]. Accordingly, intrinsic motivation means that a person is motivated because they believe in the values underlying something or because they find something stimulating and enjoyable. On the other hand, extrinsic motivation refers to the expectation of reward or money in return for one's actions. According to self-determination theory (SDT) [19], humans are naturally proactive, have a strong intrinsic desire for growth, and have basic human needs such as competence, relatedness, and autonomy, which promote motivation. For gamification-based systems to be engaging, they are expected to activate these motivators. For example, according to Kanat et al.'s [16] gamification proposal, competence and relatedness should be positioned to allow participants to see their progress and compare it with other players. Autonomy should be positioned within gamification by allowing users to choose the projects they care about. In our study, we designed a game called "fair-test" based on the concept of gamification to ensure citizen participation in the development of tests to ensure fairness in artificial intelligence, motivate participants, and make the interaction sustainable. We used extrinsic motivation elements in the game we designed to mobilize the basic motivators. In other words, we adopted an approach that rewards participants by publicizing their achievements to the masses. In terms of fairness, in our design, each study that finds any problematic element in the text and proves its success to a certain extent by any method

it wishes is accepted as a test case. The cumulative sum of these test cases is called a test suite. Each test suite tests an atomic structure. These atomic structures test the fairness of texts in terms of legal laws.

In general, everyone who cares about and works in artificial intelligence fairness constitutes the target audience of our design. In particular, software developers who produce solutions on fairness constitute the participants of the fair-test game, while everyone affected by artificial intelligence fairness constitutes an end user. Our design is expected to become a ground where current practices on fairness can be applied and contribute to the sustainability of contributions on this issue by gathering them on a single platform. In future studies, it is planned to realize a plot application of this structure designed with the gamification method.

In the remainder of this study, previous research on the subject is mentioned, then the gamification and gamification elements are explained. Next, the study's methodology is presented and the design is tested through a sample scenario. Finally, the outcomes of the scenario are discussed and the conclusion of our research are presented.

2. Related Work

In this section, because our study aims to bring together the scattered proposals for AI fairness and to develop and automate these proposals, the current literature on determining AI fairness and general proposals for eliminating fairness problems are included. Second, gamification and AI-focused research are examined, as gamification elements are used to improve AI fairness in our study.

2.1. Artificial Intelligence Fairness

Artificial intelligence fairness is a highly topical and increasingly popular topic. Chouldechova and Roth have shown that decision support systems can inadvertently encode existing human biases and introduce new ones [20]. Another study found that women are stereotyped and systematically underrepresented when the results of a visual search for doctors or nurses are compared to the actual percentage estimated by the US Bureau of Labor and Statistics [21]. The first of the two interesting results from the study is that people are more likely to prefer and rate search results that are consistent with stereotypes. The second interesting result is that when the representation of gender in visual search results is changed, people's perceptions of the real-world distribution tend to change as well. Another well-known example is the COMPAS system, a commercial tool that uses a risk assessment algorithm to predict certain categories of future crime. Specifically, this tool is used in courts in the US to assist with bail and sentencing decisions. It has been found that the false positive rate, i.e., people labeled as high risk by the tool who do not re-offend, is almost twice as high for African-Americans than for White defendants [22]. This often means that the ubiquitous use of decision support systems can create potential threats such as economic loss, social stigmatization, and even loss of liberty.

Movie and music platforms, advertisements, social media and news outlets, and search and recommendation engines are all data-driven systems [23]. Such systems are central in shaping our experiences and influencing our perception of the world. For example, how most of us listen to music now presents certain built-in bias issues. This is because when a streaming service offers music recommendations, it does so by examining what music has been listened to before. This creates a recommendation loop that reinforces existing bias and reduces diversity. A recent study analyzed the public listening records of 330,000 users of a service and found that female artists represented only 25% of the music listened to by users. The study identified gender justice as one of the main concerns of artists, as female artists do not receive equal space in music recommendations [24]. Another study by the University of Southern California on Facebook ad recommendations found that the recommendation system disproportionately showed certain job postings to men and women [25]. A recent survey has focused on fairness in these ranking [26]. Recent studies focusing on concepts and measures of fairness and the challenges involved in

applying them to recommendations and information retrieval, as well as scoring methods, have been presented by [27–29], respectively.

Pitoura et al. [10] proposed a preprocessing approach for solving fair ranking and recommendations with the aim of transforming the data to remove any underlying bias or discrimination, along with an in-processing approach to introduce new algorithms and a postprocessing approach to change the state. Chierichetti et al. [30] developed a graph clustering and ranking method. Hu and Chen [31] identified the persistence of racial inequalities and designed a number of solutions based on a dynamic reputation model of the job market, highlighting the consequences of groups' differential access to resources. In addition, there have been studies on the fairness problem in online job markets [32], on solving the social matching problem [33], and on the fairness problem in ranking and recommendation [10,34]. Most of the recent work on fairness seems to be isolated to specific tasks, focusing on the classification task for non-discrimination. However, there is a continuing need for research and solutions that consider fairness across the entire data pipeline [35].

As can be seen, there is a wide variety of research in the literature on AI fairness from many different fields. In the vast majority of these studies, query, ranking, and classification issues have been addressed through the lens of discrimination. To the best of our knowledge, no holistic approach to AI fairness has been developed. Moreover, there are no studies on the use of gamification practices in solving the fairness problem. Unlike the literature, in our study, in order to overcome the difficulties involved in automating fairness in texts, a new fairness testing paradigm using gamification is developed to gather disparate proposals about fairness on a single platform, then test them in order to develop the most effective method, thereby ensuring that the effort is sustainable.

2.2. Gamification and Artificial Intelligence

In the literature, artificial intelligence research using gamification elements has proliferated in recent years. In a recent study on AI and gamification, a machine learning algorithm-based personalized content selection for gamified systems has been proposed. The process is based on Deterding's 2015 gamified design framework, the lens of intrinsic skill atoms, and heuristics for effective design [36]. In the work of [36], machine learning was used to realize gamification. In our work, gamification is used to improve AI fairness.

Based on the fact that personalized adaptive gamification can increase individuals' motivation and performance, [37] presented a machine learning method that uses task information and individual facial expression data. Another study proposed a machine learning algorithm-based personalized content selection for gamified systems [38]. The process is based on Deterding's 2015 gamified design framework, the intrinsic skill atoms lens, and heuristics for effective design. This process has been demonstrated through the application of personalized gamification for a computer-supported collaborative learning environment. In the studies of [37,38], the authors used machine learning to provide direct gamification or demonstrate gamification results. Unlike these previous studies, in our study the gamification elements are used in the context of artificial intelligence fairness.

Another gamification-focused study examined the impact of social learning and gamification methodologies on learning outcomes in higher education. In this framework, students were asked to design, execute, and evaluate a series of learning tasks and games in two consecutive semesters of an undergraduate course [39]. Unlike our study, their [39] study focused on the effects of gamification on education and learning. In another study in the field of education, a design was developed to increase students' motivation by using a ranking of engagement levels and incorporating gamification to allow students to reinforce learning whenever and wherever they wanted to [40]. In another study, an adaptive gamified learning system (AGLS) was developed combining gamification, classification, and adaptation techniques to improve the effectiveness of e-learning. The results showed that AGLS positively impacted students' engagement and learning performance compared to gamification alone [41]. In a different study, the scope and future challenges of such

applications were examined alongside a review of the existing literature on adaptive gamification in e-learning [42]. Although all of these studies were conducted in the education field, they have in common with our study that they utilized elements of "gamification for a purpose".

In a different study, an approach using a blockchain-based solution enhanced with a gamification component for fake media or Internet of fake media objects (IoFMT) was proposed [43]. In another study, which sought to open up new ways to increase attractiveness using artificial intelligence and machine learning, an open-ended gamification method was proposed that did not depend on the course and program studied [44]. Another study argued that games can enhance the possibilities for democratic deliberation during government consultation with the public. Key design features developed in this context include game origin, management, and oversight; networked small groups at the center of the project; and artificial intelligence and automated metrics to measure deliberation [45]. Finally, another study presented a visual analytics technique for interactive data labeling that applies gamification and explainable AI concepts to support complex classification tasks [46]. Several studies by Chen (2020) [44–46] utilized elements of "gamification for a specific purpose", albeit with completely different topics and objectives than our study.

To summarize, unlike the previous studies in the literature, in this study gamification elements are utilized to gather studies on ensuring AI fairness onto a single platform, allowing AI fairness to be developed faster, more efficiently, and through a more sustainable process.

3. Method

In this study, a game design was developed to enable civic engagement around the development of AI fairness tests and to bring together scattered proposals on a single platform. The design we developed consists of both Rule and Discussion (Zero-Suppressed Decision Diagram [47], Centrality) elements and Test (Confusion Matrix, Mutation Test) elements. The Rule and Discussion elements form the Zero-Suppressed Decision Diagram (ZDD) and Centrality scores in the design we developed. In contrast, the Test element forms the Confusion Matrix and Mutation Test scores.

Figure 1 shows the general design of the gamification elements discussed in detail in this section.

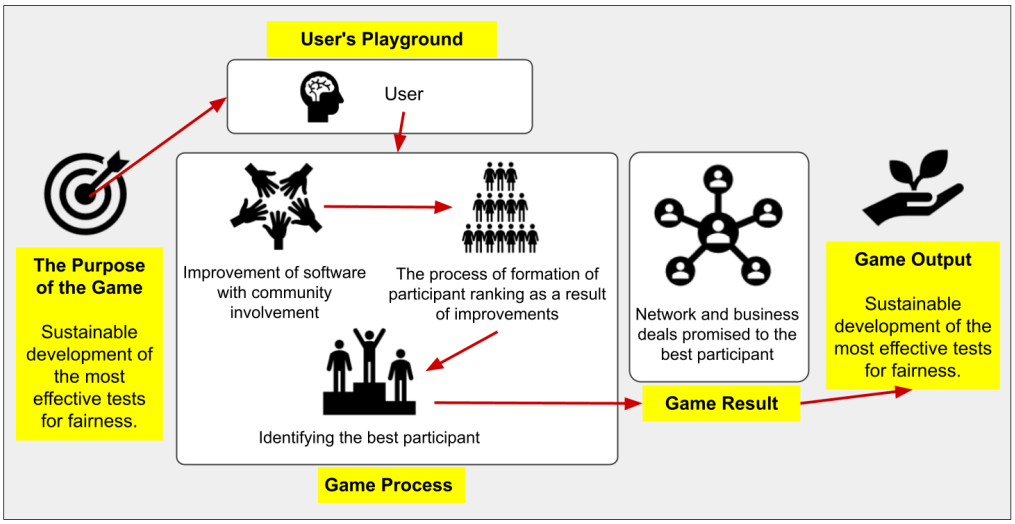


Figure 1. General design of gamification.

3.1. Gamification Elements

Gamification elements are considered the most essential design components of a gamification concept [48]. In gamification, it is essential to understand the role of existing

elements, what they represent, and their properties [49]. In this section, the elements of the design and their roles are explained.

3.1.1. Rule and Comment

Rules are structures used to test whether a text is fair. All the rules used form a rule network. This network contains only rules and comments. Here, rules are immutable, while comments are mutable. Users set various rules and comments in the system. The popularity of these rules and comments is calculated by a Zero-Suppressed Decision Diagram (ZDD) and a Centrality score.

Each rule is treated as a node in the graph structure of the system. There are three types of relationships in the rule network. Logical relationships between rules in the network are found in three different forms: verification (AND), refutation (NOT), and alternative generation (OR).

For example, the sentence "women and men are equal" is a rule. If desired, this rule can be supported by the rule "women have the same working conditions as men" (AND). Propositions that support, refute, or provide an alternative point of view, such as "women and men are not equal" (NOT) or "women and men have very different qualities and are therefore different" (OR), are expressed as rules.

The score of these relationships is calculated with the Zero-Suppressed Decision Diagram (ZDD), as it has a logical structure within a domain that users can specify. In addition, we use calculations based on centrality criteria in the network to determine the popularity of rules and comments. The reason for using ZDD is that it can express rules and discussions canonically as a "sum of products". In this way, supported and unsupported rules are kept compact and easy to find.

As shown in Figure 2, Rule A has associations. The score of shared rules increases or decreases depending on how much that rule is verified in the rule network. The individual score of the person sharing a rule affects the score of the rule they share to a certain extent. This means that rules shared by people with a high game score have more influence.

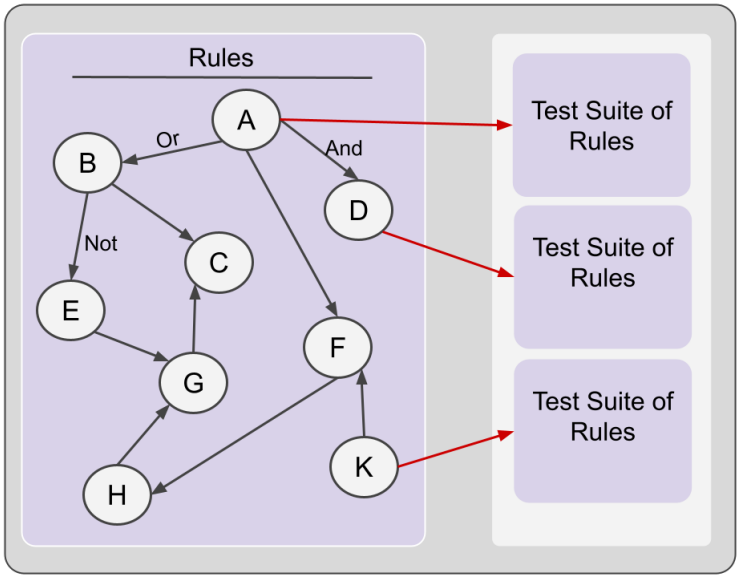


Figure 2. Rule A's relationships.

Another criterion used in the generation of rules and their interpretation is centrality. Centrality is a computational method used in graph theory and network analysis to rank nodes according to their position in the network [50]. The aim is to understand the density of a node according to specific criteria. There are many measures of centrality, including degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality.

Centrality values are used in this design to determine the importance of the rules developed within the game. Accordingly, rules with higher centrality values are more important. Centrality metrics are included in the system as a score related to the node's importance. This score represents the centrality score of the rule.

3.1.2. Text Testing

This is a method for testing the conformity of texts to rules. Users can add new rules or tests to the system. Each rule can have subrules. Here, tests are prepared to determine the conformity of the text to the rule, and the text is tested through these tests. Each subrule can be tested independently, and each subrule must pass the tests of its parent rules.

Test cases are test procedures that can be run successfully or unsuccessfully. Negative test cases should produce negative results, while positive ones should yield positive results. A test is successful if the expected results match the actual results [51].

A test suite is a set of test cases [52]. Here, the output of one test case can be the input of another test case.

As can be seen in Figure 2, each rule in the rule network has a test suite. In our design, the performance of the test cases is measured by the confusion matrix and the performance of the test suites is measured by mutation testing.

In our design, test cases can be developed with simple natural language processing methods or with more complex structures such as machine learning [53]. The Confusion Matrix measures the performance of the generated test cases. Accuracy, Precision, Recall, and F1 Score [54] are calculated based on the Confusion Matrix.

Another metric used in the testing phase is mutation testing. Mutation testing is a method used to measure the performance of test suites [55]. Although it has standard features with the confusion matrix method, it differs in its holistic approach.

In this study, mutants are treated as texts in mutation testing. The text added to the system, i.e., the mutant, can be reproduced with the mutation operator. If the test case captures a mutant, this is called a dead mutant. If it is missed, it is called a living mutant. The performance of the test suite is tested using these mutations. Figure 3 shows the test suite's test cases and each test case's mutants. In our model, each test case can have many mutants.

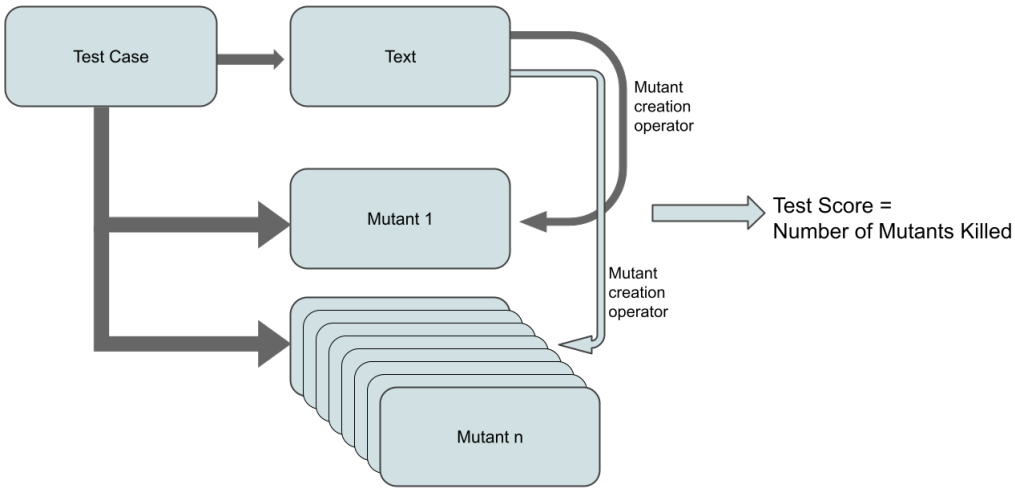


Figure 3. General representation of mutants.

In mutation testing, various metrics are used to test the test suite. The most commonly used ones are the Fault Detection Ratio (FDR), Average Percentage of Faults Detected (APFD), FDR/Test Suite Size, FDR/Test Case Time, and FDR/Test Suite Time [52]. The FDR formula has the same meaning as Accuracy in our design. Apart from this, APFD is preferred for use in cases with multiple test cases, as it provides information about

how to arrange the test cases in cases with many test cases. FDR/Test Suite Size provides information in terms of the test suite. It differs from the confusion matrix in this respect. On the other hand, the confusion matrix does not produce any output related to timing. In contrast, FDR/Test Case Time and FDR/Test Suite Time provide information about timing. The performance of the test suite is determined using these scores. The higher the performance of the test suite, the higher the detection rate of the tested rule in the text. At the same time, the overall performance of the test suite constitutes the test score.

3.1.3. Score

There are two basic game scores in the design, the rule score and the test score; their sum determines the user score. The rule score is obtained by normalizing the ZDD score and the centrality score to a range of 0–1, while the test score is obtained by normalizing the confusion matrix score and the mutation test score to the range of 0–1.

$$Rule\,score = ZDD\,score + Centrality\,score \tag{1}$$

APFD generates values between 0 and 1, while FDR/Test Suite Size and FDR/Test Suite Time generate non-normalized values. Therefore, the FDR/Test Suite Size and FDR/Test Suite Time values are normalized and added to the formula. In our model, the formula used for the test score is as follows:

$$Mutant\,score = APFD + FDR/Test\,Suite\,Size + FDR/Test\,Suite\,Time \tag{2}$$

$$Test\,Score = Mutant\,Score + Confusion\,Matrix\,Score \tag{3}$$

The user score is the sum of the rule score and the test score normalized to a range of 0–0.5:

$$User\,Score = Rule\,score + Test\,score \tag{4}$$

Figure 4 shows an overview of the user's playing field and the elements that make up the user score. Here, the user score, which determines the user's influence factor in the game, is composed of the rule score and the test score.

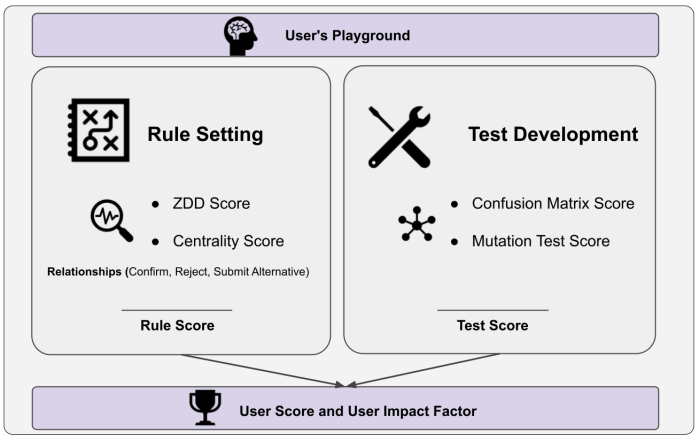


Figure 4. The user's playground and the elements that make up the user score.

3.2. Player Action Steps

When the user logs into the game area, they can perform various action steps. Accordingly, the user enters rules and comments into the game system. The rules and comments entered by the user create the rule score. The user's rule score is determined according to the centrality and ZDD scores. The user can enter a test case, mutant operator, and mutant into the game system. The test case score entered by the user is calculated using the mutation test and the confusion matrix.

As can be seen in Table 1, each rule, comment, test, mutant operator, and mutant entered into the system determines the user's rule score and test score. The rule and test scores generated in this way determine the user's score in the game. The user score constitutes the user's influence factor in the game system, and ensures that each rule, comment, test, mutant operator, or mutant that players with a high user score enter into the system has a greater influence in the game. The user's action steps in the game are shown in Figure 5.

Table 1. Summary of game metrics.

Game Element	Criteria Used	Description	Scoring
Rule and Comment	Zero-Suppressed Decision Diagram Score	Allows parsing and scoring relational rules and comments. High processing power is required	Rule and Comment Score = ZDD score + Centrality score
	Centrality Score	Allows rules and comments to be scored without parsing them	
Test	Confusion Matrix Score	Generates a test performance score using Accuracy, Precision, Recall, and F1 Score	Test Score = Confusion Matrix Score + Mutation Test Score
	Mutation Test Score	Generates a test performance score using the Fault Detection Ratio (FDR), Average Percentage of Faults Detected (APFD) FDR/Test Suite Size, FDR/Test Case Time, and FDR/Test Suite Time	

To illustrate the example scenario, in Figure 5, A1, A2, and A3. . . denote the actors (users) logging into the system, while H1, H2, H3. . . denote the rules circulating in the system.

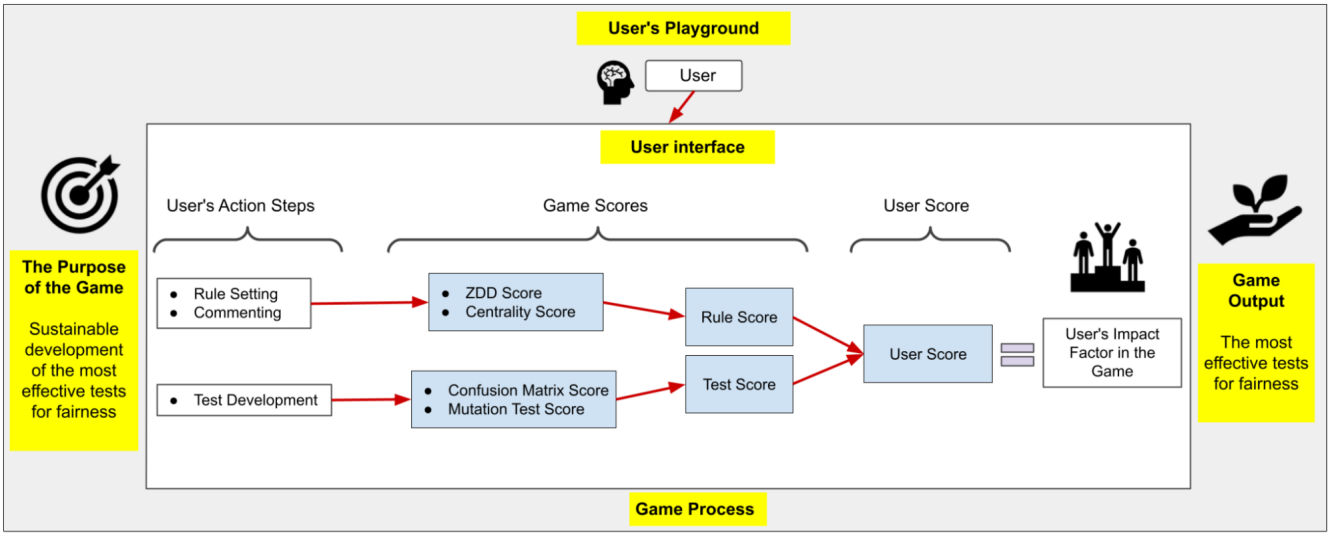


Figure 5. Overview of the user's playground and action steps.

Including rules introduced by different users creates different versions of that rule in the network, and all of these versions are stored in the rule. On the other hand, a rule's score is calculated based on all logical relations related to the rule stored in the ZDD. For example, the ZDD representation and score of Rule H1 are shown in Figure 5.

Accordingly, the score of Rule H1 within a user-specified domain is the sum of the Rules H1 and H2 that validate this rule, i.e., 2. In addition to this, as can be seen, for example, in Figure 6, the structure we designed contains many different scores that can be evaluated within the framework of graph centralities such as Degree, Eigenvector, and Closeness, which are calculated and shared with the users in order to follow the relationships and patterns related to the rules. As can be seen in Figure 6, when user A1 shares rule H1 in the network, rules H2 and H3 validate rule H1. The red arrows show the verification relationship between the rules in Figure 6. The black arrows indicate comment relations.

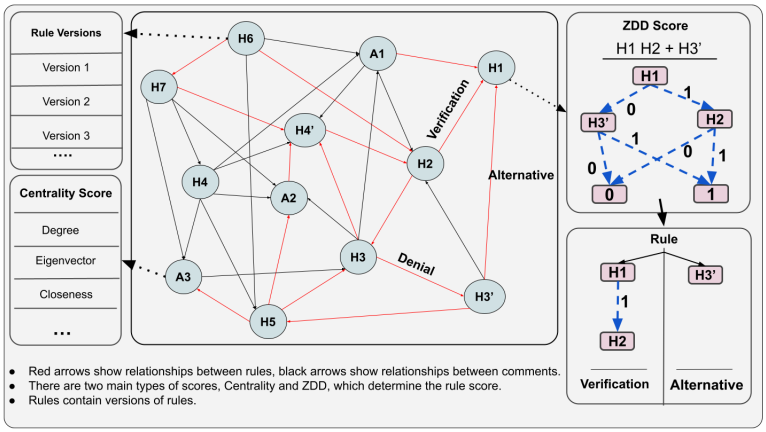


Figure 6. General structure of the rule network.

Thanks to its customizable interface, all participants can log in with their accounts, and each participant has a rule-sharing profile in the system. Thanks to this personal profile, participants can follow every event related to their rules and tests from their accounts. When a user logs in to the system, they can see all the rules and tests they have created, shared, or followed, along with other rules and difficulties related to the rule or test, summarized on their screen. Users can manage rules and tests related to them on their profile page. The user can interpret any rule and confirm the accuracy or inaccuracy of their rules and tests. Users can track which rules and tests are generally popular when logging into the system. They can comment on issues related to their own rules and tests, create new rules and tests, and delete or modify the rules and tests they have created. On the other hand, it can be considered a limitation that malicious users may include misleading tests about fairness. However, adding new mutants (texts) to the system should eliminate this problem in the medium and long term, as a user who creates such a misleading situation will be penalized or banned from the system. When the system starts up, there will need to be more participants, meaning that performance will be low. To overcome this problem, we plan to focus on promotional activities. In addition, it can be considered that the game design we have developed is not very rich in reward at this stage. However, it is envisaged that reward mechanisms will be developed over time, as this becomes known and seen as an area of attraction by companies with the effect of promotional activities.

4. Sample Scenario

An actor enters the system and introduces Rule A: "men and women are equal". A different set of rules associated with that rule, brought by different actors, are then integrated into the system. Figure 7 illustrates this structure.

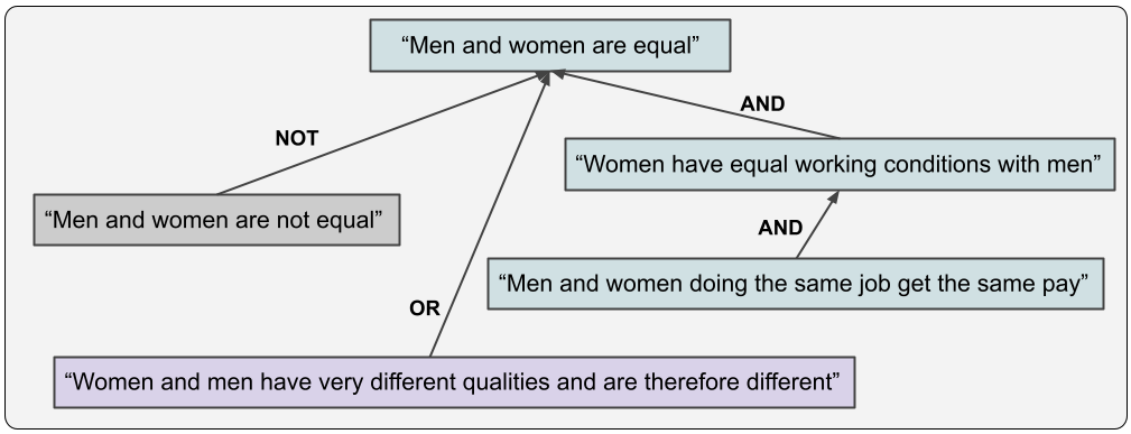


Figure 7. Example scenario.

As shown in Figure 7, Rule A, "Men and women are equal", is a rule. If desired, this rule can be supported by Rule B: "Women have equal working conditions with men" (AND), or C: "Women and men are not equal" (NOT) and D: "Women and men are paid the same for doing the same work", or E: "Women and men have very different qualities and are therefore different" (OR). With respect to the original rule, these variously support it, refute it, or provide an alternative point of view. The ZDD representation of this structure is described in detail below.

In Figure 8, because Rule B is negative, $4 - 1 = 3$ represents the number of rules favoring Rule A. On the other hand, according to the same scenario, the Degree Centrality value of Rule A is 2. The Betweenness Centrality value of Rule A is 4/8, i.e., 0.5, because it is located on the paths AA, AC, AD, BC, BD, CD, AE, and DD among the paths AB, AC, AD, BC, BD, CD, AE, and DD. The Closeness Centrality is 1/7, i.e., 0.14, because AB = 1, AC = 2, AD = 3, and AE = 1.

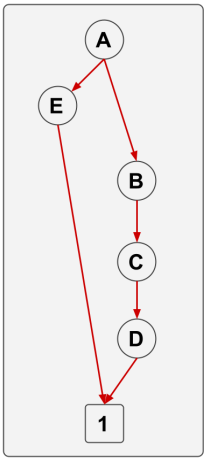


Figure 8. ZDD representation of the example scenario.

The total score of Rule A's related rules and interpretations, i.e., the unnormalized sum of the ZDD and centrality scores, is 3 + 2 + 0.5 + 0.14 = 5.64.

Suppose another user adds a test case to Rule A. Various other users then add mutants to this test case. For example, let us say that ten mutants are created. The confusion matrix values of Rule A in this case are shown in Table 2.

Table 2. Confusion matrix for Rule A.

		Predicted
Actual	3	1
	2	4

For example, in this case, the Accuracy value of Rule A is 7/10, the Precision value is 3/5, the Recall value is 3/4, and the F1 score is 0.45/1.35 = 0.75. Thus, the Confusion Matrix score is 0.75.

When we calculate the mutation test score of Rule A, FDR = 7/10. Because there is one test case, the APFD value is 1. The FDR/Test Suite Size score is equal to FDR. Assuming that the test case can compute a mutant every ten seconds, the Test Suite Time is 0.07 from FDR/10. Because there is one test case, the Test Case Time equals the Test Suite Time.

Mutant score = 1 + 0.07 + 0.07 + 0.07 = 1.14; Test Score = 1.14 + 0.75 = 1.89.

The score of the user who made Rule A = Rule score + Test score = 5.64 + 1.89 = 7.53.

Because normalization is not used in the above calculations, an approximate sample result is obtained. Furthermore, when a user adds a test case to a rule, the success score of the test case is added to that person's user score. The user score of the person who generates mutants to the test case is increased by the number of live mutants.

5. Conclusions

In this study, gathering disparate proposals on fairness on a single platform was taken as a starting point based on the difficulties in automating fairness in texts. In this framework, a new fairness testing paradigm called "fair-test", which uses gamification elements, was developed to use with proposals on AI fairness in order to test these proposals and find the most effective method, thereby strengthening civic participation in AI fairness efforts and making them sustainable. In this context, a game was designed that allows participants to see their progress and compare it with other players. It uses extrinsic motivation at its core, i.e., rewarding participants by publicizing their achievements to the masses.

The game design we developed consists of Rules and Comments (ZDD, Centrality) and Test (Confusion Matrix, Mutation Test) elements that are created to determine the fairness in texts. In this design, the Rule and Discussion elements constitute the ZDD and Centrality scores, while the Test element constitutes the Confusion Matrix and Mutation Test scores. The sum of the Rule and Discussion and Test scores, which are the basic gamification elements, determines the user score, and this score constitutes the user's influence factor in the game. In our design, rules, comments, and tests can be developed with simple natural language processing methods as well as more complex and different structures such as machine learning. This provides participants with the flexibility to develop tests with a wide variety of methods while contributing to AI fairness, which is the study's main goal.

This model utilizes the social network effect to prevent unfair textual data by focusing on "sharing, discussion, and verifiability of generated rules and tests" to prevent unfair textual data. It is expected that our design will become a ground where current practices on fairness can be applied, mediating the sustainability of contributions on this issue and thereby contributing to the current orientation on fairness. In future studies, it is planned to realize a plot application of this structure designed with the gamification method.

Author Contributions: Conceptualization, S.T. and D.E.; methodology, S.T.; software, S.T.; validation, S.T., D.E.; formal analysis, S.T.; investigation, D.E.; resources, S.T.; writing—original draft preparation, S.T.; writing—review and editing, D.E. and G.K.; visualization, D.E.; supervision, S.T.; project administration, S.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data is unavailable in this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kleinberg, J.; Mullainathan, S.; Raghavan, M. Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv* **2016**, arXiv:1609.05807.
2. Grgic-Hlaca, N.; Redmiles, E.M.; Gummadi, K.P.; Weller, A. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In Proceedings of the 2018 World Wide Web Conference, International World Wide Web Conferences Steering Committee, Lyon, France, 23–27 April 2018; pp. 903–912.
3. Plane, A.C.; Redmiles, E.M.; Mazurek, M.L.; Tschantz, M.C. Exploring User Perceptions of Discrimination in Online Targeted Advertising. In Proceedings of the 26th USENIX Security Symposium (USENIX Security 17), Vancouver, BC, Canada, 16–18 August 2017; pp. 935–951.
4. Makhlouf, K.; Zhioua, S.; Palamidessi, C. Machine learning fairness notions: Bridging the gap with real-world applications. *Inf. Process. Manag.* **2021**, *58*, 102642. [CrossRef]
5. Tian, H.; Zhu, T.; Liu, W.; Zhou, W. Image fairness in deep learning: Problems, models, and challenges. *Neural Comput. Appl.* **2022**, *34*, 12875–12893. [CrossRef]
6. Asudeh, A.; Jin, Z.; Jagadish, H.V. Assessing and Remedying Coverage for a Given Dataset. In Proceedings of the 2019 IEEE 35th International Conference on Data Engineering (ICDE), Macao, China, 8–12 April 2019; pp. 554–565.
7. Friedler, S.A.; Scheidegger, C.; Venkatasubramanian, S. On the (im)possibility of fairness. *arXiv* **2016**, arXiv:1609.07236.

8. Friedler, S.A.; Scheidegger, C.; Venkatasubramanian, S. The (Im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Commun. ACM* **2021**, *64*, 136–143. [CrossRef]
9. Pitoura, E. Social-minded Measures of Data Quality: Fairness, Diversity, and Lack of Bias. *J. Data Inf. Qual.* **2020**, *12*, 1–8. [CrossRef]
10. Pitoura, E.; Stefanidis, K.; Koutrika, G. Fairness in rankings and recommendations: An overview. *VLDB J.* **2022**, *31*, 431–458. [CrossRef]
11. Kim, B.; Park, J.; Suh, J. Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information. *Decis. Support Syst.* **2020**, *134*, 113302. [CrossRef]
12. Hamari, J. Gamification. In *The Blackwell Encyclopedia of Sociology*; Wiley: Hoboken, NJ, USA, 2019; pp. 1–3.
13. Jennings, J. Urban Planning, Community Participation, and the Roxbury Master Plan in Boston. *Ann. Am. Acad. Political Soc. Sci.* **2004**, *594*, 12–33. [CrossRef]
14. Pelling, N. The (short) prehistory of gamification, Funding Startups (& other impossibilities). *J. Nano Dome* **2011**. Available online: https://nanodome.wordpress.com/2011/08/09/the-short-prehistory-of-gamification/ (accessed on 16 September 2022).
15. Deterding, S.; Dixon, D.; Khaled, R.; Nacke, L. From game design elements to gamefulness. In Proceedings of the 15th International Academic MindTrek Conference on Envisioning Future Media Environments—MindTrek '11, Tampere, Finland, 28–30 September 2011; ACM Press: New York, NY, USA, 2011.
16. Kanat, I.E.; Siloju, S.; Raghu, T.S.; Vinze, A.S. Gamification of emergency response training: A public health example. In Proceedings of the 2013 IEEE International Conference on Intelligence and Security Informatics, Seattle, WA, USA, 4–7 June 2013; pp. 134–136.
17. Romano, M.; Díaz, P.; Aedo, I. Gamification-less: May gamification really foster civic participation? A controlled field experiment. *J. Ambient Intell. Humaniz. Comput.* **2022**, *13*, 4451–4465. [CrossRef]
18. Malone, T.W. Toward a theory of intrinsically motivating instruction. *Cogn. Sci.* **1981**, *5*, 333–369. [CrossRef]
19. Ryan, R.M.; Deci, E.L. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am. Psychol.* **2000**, *55*, 68–78. [CrossRef]
20. Chouldechova, A.; Roth, A. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* **2020**, *63*, 82–89. [CrossRef]
21. Kay, M.; Matuszek, C.; Munson, S.A. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Seoul, Republic of Korea, 18–23 April 2015; Association for Computing Machinery: New York, NY, USA, 2015; pp. 3819–3828.
22. Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. Machine Bias. Propublica. 2016. Available online: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed on 23 September 2022).
23. Martin, N. How Social Media Has Changed How We Consume News. *Forbes Magazine*, 13 November 2018.
24. Ferraro, A.; Serra, X.; Bauer, C. Break the Loop: Gender Imbalance in Music Recommenders. In Proceedings of the 2021 Conference on Human Information Interaction and Retrieval, Online, 14–19 March 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 249–254.
25. Imana, B.; Korolova, A.; Heidemann, J. Auditing for Discrimination in Algorithms Delivering Job Ads. In Proceedings of the Web Conference 2021, New York, NY, USA, 19–23 April 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 3767–3778.
26. Zehlike, M.; Yang, K.; Stoyanovich, J. Fairness in Ranking: A Survey. *arXiv* **2021**, arXiv:2103.14000.
27. Ekstrand, M.D.; Burke, R.; Diaz, F. Fairness and Discrimination in Retrieval and Recommendation. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 1403–1404.
28. Asudeh, A.; Jagadish, H.V. Fairly evaluating and scoring items in a data set. *Proc. VLDB Endow.* **2020**, *13*, 3445–3448. [CrossRef]
29. Oosterhuis, H.; Jagerman, R.; de Rijke, M. Unbiased Learning to Rank: Counterfactual and Online Approaches. In Proceedings of the Web Conference 2020, Taibei, Taiwan, 20–24 April 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 299–300.
30. Chierichetti, F.; Kumar, R.; Lattanzi, S.; Vassilvitskii, S. Fair Clustering Through Fairlets. *arXiv* **2018**, arXiv:1802.05733.
31. Hu, L.; Chen, Y. A Short-term Intervention for Long-term Fairness in the Labor Market. In Proceedings of the 2018 World Wide Web Conference, International World Wide Web Conferences Steering Committee: Republic and Canton of Geneva, CHE, Lyon, France, 23–27 April 2018; pp. 1389–1398.
32. Elbassuoni, S.; Amer-Yahia, S.; Ghizzawi, A. Fairness of Scoring in Online Job Marketplaces. *ACM/IMS Trans. Data Sci.* **2020**, *1*, 1–30. [CrossRef]
33. Olsson, T.; Huhtamäki, J.; Kärkkäinen, H. Directions for professional social matching systems. *Commun. ACM* **2020**, *63*, 60–69. [CrossRef]
34. Machado, L.; Stefanidis, K. Fair Team Recommendations for Multidisciplinary Projects. In Proceedings of the 2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI), Thessaloniki, Greece, 14–17 October 2019; pp. 293–297.
35. Stoyanovich, J.; Howe, B.; Abiteboul, S.; Miklau, G.; Sahuguet, A.; Weikum, G. Fides: Towards a Platform for Responsible Data Science. In Proceedings of the 29th International Conference on Scientific and Statistical Database Management, Chicago, IL, USA, 27–29 June 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 1–6.

36. Bucea-Manea-Tonis, R.; Kuleto, V.; Gudei, S.C.D.; Lianu, C.; Lianu, C.; Ilic, M.P.; Paun, D. Artificial Intelligence Potential in Higher Education Institutions Enhanced Learning Environment in Romania and Serbia. *Sustain. Sci. Pract. Policy* **2022**, *14*, 5842. [CrossRef]
37. Lopez, C.; Tucker, C. Toward Personalized Adaptive Gamification: A Machine Learning Model for Predicting Performance. *IEEE Trans. Comput. Intell. AI Games* **2020**, *12*, 155–168. [CrossRef]
38. Knutas, A.; van Roy, R.; Hynninen, T.; Granato, M.; Kasurinen, J.; Ikonen, J. A process for designing algorithm-based personalized gamification. *Multimed. Tools Appl.* **2019**, *78*, 13593–13612. [CrossRef]
39. Sousa-Vieira, M.E.; Lopez-Ardao, J.C.; Fernandez-Veiga, M.; Rodriguez-Rubio, R.F. Study of the impact of social learning and gamification methodologies on learning results in higher education. *Comput. Appl. Eng. Educ.* **2022**, *31*, 131–153. [CrossRef]
40. Jimenez-Hernandez, E.M.; Oktaba, H.; Diaz-Barriga, F.; Piattini, M. Using web-based gamified software to learn Boolean algebra simplification in a blended learning setting. *Comput. Appl. Eng. Educ.* **2020**, *28*, 1591–1611. [CrossRef]
41. Daghestani, L.F.; Ibrahim, L.F.; Al-Towirgi, R.S.; Salman, H.A. Adapting gamified learning systems using educational data mining techniques. *Comput. Appl. Eng. Educ.* **2020**, *28*, 568–589. [CrossRef]
42. Bennani, S.; Maalel, A.; Ben Ghezala, H. Adaptive gamification in E-learning: A literature review and future challenges. *Comput. Appl. Eng. Educ.* **2022**, *30*, 628–642. [CrossRef]
43. Chen, Q.; Srivastava, G.; Parizi, R.M.; Aloqaily, M.; Al Ridhawi, I. An incentive-aware blockchain-based solution for internet of fake media things. *Inf. Process. Manag.* **2020**, *57*, 102370. [CrossRef]
44. Duggal, K.; Gupta, L.R.; Singh, P. Gamification and Machine Learning Inspired Approach for Classroom Engagement and Learning. *Math. Probl. Eng.* **2021**, *2021*, 9922775. [CrossRef]
45. Gastil, J. To Play Is the Thing: How Game Design Principles Can Make Online Deliberation Compelling. *Am. Behav. Sci.* **2022**. [CrossRef]
46. Sevastjanova, R.; Jentner, W.; Sperrle, F.; Kehlbeck, R.; Bernard, J.; El-assady, M. QuestionComb: A Gamification Approach for the Visual Explanation of Linguistic Phenomena through Interactive Labeling. *ACM Trans. Interact. Intell. Syst.* **2021**, *11*, 19. [CrossRef]
47. Minato, S.I. *Binary Decision Diagrams and Applications for VLSI CAD*; Springer: Berlin/Heidelberg, Germany, 2011.
48. Schöbel, S.M.; Janson, A.; Söllner, M. Capturing the complexity of gamification elements: A holistic approach for analysing existing and deriving novel gamification designs. *Eur. J. Inf. Syst.* **2020**, *29*, 641–668. [CrossRef]
49. Liu, D.; Santhanam, R.; Webster, J. Toward meaningful engagement: A framework for design and research of gamified information systems. *MIS Q.* **2017**, *41*, 1011–1034. [CrossRef]
50. Freeman, L.C. Centrality in social networks conceptual clarification. *Soc. Netw.* **1978**, *1*, 215–239. [CrossRef]
51. Cangussu, J.W.; DeCarlo, R.A.; Mathur, A.P.; . A formal model of the software test process. *IEEE Trans. Softw. Eng.* **2002**, *28*, 782–796. [CrossRef]
52. Mathur, A. *Foundations of Software Testing*, 2nd ed.; Pearson: Delhi, India, 2013.
53. Chalkidis, I.; Kampas, D. Deep learning in law: Early adaptation and legal word embeddings trained on large corpora. *Artif. Intell. Law* **2019**, *27*, 171–198. [CrossRef]
54. Alpaydin, E. *Introduction to Machine Learning*, 4th ed.; MIT Press: Cambridge, MA, USA, 2020. Available online: https://www.google.com/books?hl=tr&lr=&id=tZnSDwAAQBAJ&oi=fnd&pg=PR7&dq=ethem+alpayd%C4%B1n&ots=F3YR7UdwBg&sig=yjp6CpKWhkc2puDmpK4tsoD-X5I (accessed on 28 December 2022).
55. Wong, W.E.; Mathur, A.P. Reducing the cost of mutation testing: An empirical study. *J. Syst. Softw.* **1995**, *31*, 185–196. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.