Related papers at https://media-bias-research.org/publications/

Preprint of the paper:

Hinterreiter, S. & Spinde, T. & Oberdörfer, S. & Echizen, I. & Latoschik, M. E., "News Ninja: Gamified Annotation of Linguistic Bias in Online News", in Proc. ACM Hum.-Comput. Interact., Vol. 8, No. CHI PLAY, Article 327, DOI: 10.1145/3677092.

Click to download: BibTeX

News Ninja: Gamified Annotation of Linguistic Bias in Online News

SMI HINTERREITER, University of Würzburg, Germany
TIMO SPINDE, University of Göttingen, Germany
SEBASTIAN OBERDÖRFER, University of Würzburg, Germany
ISAO ECHIZEN, National Institute of Informatics, Japan
MARC ERICH LATOSCHIK, University of Würzburg, Germany

Fig. 1. From left to right, the first screen shows News Ninja's home screen with player metrics, the group mission, the breaking news tile, and four game modes. The second screen shows feedback on an incorrectly annotated sentence in the *Publish* game mode (Section 3.4.1), with one correct word, one incorrect word, and one missed word. The third screen displays the *Critique* mode with the same annotated sentence. Players can agree or disagree by swiping or using the buttons. The last screen shows the *Paper* section where played sentences are archived. Sentences missing a ground truth during play show up white when players can collect feedback.

Recent research shows that visualizing linguistic bias mitigates its negative effects. However, reliable automatic detection methods to generate such visualizations require costly, knowledge-intensive training data. To facilitate data collection for media bias datasets, we present News Ninja, a game employing data-collecting game mechanics to generate a crowdsourced dataset. Before annotating sentences, players are educated on media bias via a tutorial. Our findings show that datasets gathered with crowdsourced workers trained on News Ninja can reach significantly higher inter-annotator agreements than expert and crowdsourced datasets with similar data quality. As News Ninja encourages continuous play, it allows datasets to adapt to the reception and contextualization of news over time, presenting a promising strategy to reduce data collection expenses, educate players, and promote long-term bias mitigation.

CCS Concepts: • **Human-centered computing → Interactive systems and tools**; **Human computer interaction (HCI)**.

Authors' Contact Information: Smi Hinterreiter, University of Würzburg, Würzburg, Germany; Timo Spinde, University of Göttingen, Göttingen, Germany; Sebastian Oberdörfer, University of Würzburg, Würzburg, Germany; Isao Echizen, National Institute of Informatics, Tokyo, Japan; Marc Erich Latoschik, University of Würzburg, Würzburg, Germany.

Additional Key Words and Phrases: media bias, news bias, linguistic bias, crowdsourcing, Game With A Purpose

1 Introduction

Online news is the predominant information source for current events and critical political issues [1, 2, 3] and is generally perceived as trustworthy [4]. However, news media frequently carry inherent biases, a structural defect [5] that can influence public opinion [6, 7, 8]. As media bias is a multifaceted concept with various subtypes [9], we focus on linguistic bias. Linguistic bias is evident in word choice and framing [10, 11, 12] and describes the usage of language to convey a certain view of events, groups, or individuals [9]. Readers are often oblivious to this bias, which can lead to a compromised understanding of issues and promote a skewed perspective [13, 14, 15]. One effective countermeasure against the adverse effects of linguistic bias is highlighting its presence to readers [13, 16]. While possible, manual annotation of bias by experts is impractical given the sheer volume of news content [17]. Recent advancements in natural language processing (NLP) offer promise for automated bias detection [18, 19, 20]. Yet, their current performance still falls short of the required performance for end-user solutions [19]. This primarily arises from the resource-intensiveness and complexity of creating large, high-quality training datasets [21]. Biases such as spin or framing bias complicate the detection task due to their reliance on context [9]. For example, the word "killed" is unbiased in the sentence "12 people were killed." However, it becomes biased in the sentence "Trump killed his opponent in his speech."

Games With a Purpose (GWAPs) generate data and labels through human computation as a byproduct of gaming [22, 23, 24]. Concurrently, games can serve an educational function and increase awareness of the game's topic [25, 26]. Merging those concepts could increase player comprehension of the issue, thereby facilitating the collection of higher-quality data [23]. Thus, GWAPs present a viable solution to the three challenges of increasing players' bias detection skills, making annotation tasks engaging, and circumventing the need for experts for dataset creation. Specifically for a linguistic bias GWAP that requires the education of players to annotate biased words and sentences accurately, we investigate the research questions:

- (Q1) How can knowledge of linguistic bias be transferred in an interactive and gamified manner?
- (Q2) How can game mechanics facilitate the annotation task?
- (Q3) Can player-generated data achieve comparable results to expert-generated data?

This work introduces the GWAP News Ninja (Figure 1). News Ninja aims to increase players' linguistic bias detection skills before gathering data to refine automated bias detection. The design process addresses Q1 by employing game design frameworks to translate annotation guidelines [19] into game elements, including a storyline, progression elements, direct feedback, rewards, and social elements (Figure 2) [27, 28, 29, 30, 31]. For (Q2), we deconstruct the annotation process into game mechanics, which we qualitatively pre-test in combination with the tutorial (Section 4.1). Based on the results, we re-design the tutorial and label-generating game modes (Section 3.4). To assess data quality for (Q3), we develop and test News Ninja as a streamlined, responsive web application to compare player-generated data to expert-generated data (Section 4.2).

Player-generated labels show a significant 10.28% increase in inter-annotator agreement (IAA) compared to the baseline dataset BABE [19]. Experts[1] re-annotate the game sentences and achieve an agreement of 79.8%, indicating News Ninja as a promising approach for crowdsourcing expert-like linguistic bias labels. The prototype showcases the possibilities of combining game design, education, and annotation tasks. As the first functional GWAP focused on linguistic bias, News

[1]Experts are three researchers with more than one year of experience in media bias research from the university network.

Ninja introduces a GUI, tutorial, five game modes, a feedback system, and a delayed feedback mechanism for sustained player engagement when a ground truth is missing. The game can be adapted to various NLP annotation scenarios to create various linguistic datasets.

How news is perceived and contextualized is subject to continuous evolution, yet datasets remain static. Applications like News Ninja hold the potential to update datasets to mirror these changes while simultaneously enhancing public awareness and bias detection skills. We conclude by discussing our game annotation mechanics for subjective truths and potential cognitive biases within the dataset. The dataset is publicly available.[2]

2 Background

Linguistic bias is reflected in statements when language is systematically used to reinforce stereotypes or specific perceptions of groups, events, or individuals [32]. This form of bias is fine-grained, evident in individual words, phrases, or sentence structures, and can alter the context and meaning of a statement. Such bias can manifest through one-sided terms or adjectives that amplify or add subjective meaning to a sentence or text [33]. The choice of words can influence the perceived credibility of a statement [33]. Additionally, words and phrases often carry connotations, integrating subtle feelings or biases into statements [34].

2.1 Automatic Detection of Linguistic Bias

NLP classifiers show potential in algorithmically detecting and indicating linguistic bias [18, 19, 20]. The currently most common approach is fine-tuning large language models with bias datasets containing statements or sentences [9]. However, their accuracy falls short of the standards required for consumer tools [9] because extensive, high-quality bias datasets are missing [35, 9]. Due to the complex nature of the annotation task, crowdsourced datasets exhibit low agreement and higher noise [35]. Spinde et al. [35]'s crowdsourced dataset MBIC achieves a F1-score of 0.43 and an IAA of $\alpha = 0.21$. Contrasting, expert datasets are costly but achieve a higher F1-score of 0.804 and an IAA of $\alpha = 0.39$ [19]. Prior research tries to optimize cost and reliability by balancing crowdsourced and expert labels and training non-experts over extended periods to become experts [19]. However, this method remains costly for large-scale application [19], stressing the need for alternative solutions in generating reliable media bias datasets. Our strategy addresses the challenges by educating non-experts and substituting financial incentives with engaging gameplay.

2.2 Media Literacy Education and Online Civic Reasoning Approaches

We investigate interactive education methods in bias and media literacy to design game mechanics that instruct and train players [26, 36]. Educational research on media bias itself is sparse and merely focuses on the impact and the perception of media bias [37, 13, 16]. Curricula on media literacy [38] and online civic reasoning (OCR) [39, 36] touch upon bias through agenda-setting — which news ultimately gets reported — and framing — how and with which words and phrases the news is presented [40]. Generally, interventions related to News Ninja target younger demographics through school curricula [39] or employ interactive checklists [41], with interactive online courses [36] and serious games [26]. For instance, the interactive web-based learning tool "The News Evaluator" equips users with skills for critical online content engagement [36]. The application packs OCR objectives into a structured format, a tutorial, hands-on learning tasks, and implicit and explicit feedback.

[2] https://github.com/Media-Bias-Group/News-Ninja

"Bad News" is a serious game for media literacy education [26]. It teaches six misinformation strategies by embedding them into the story and letting players adopt the antagonist's role. Players build their media empire's following and credibility by generating and disseminating news, enhanced by a gameful interface, ownership elements, and achievements to foster motivation and learning. The game significantly improves players' ability to detect misinformation through inoculation by exposing them to weak doses of misinformation to build cognitive immunity [37, 26].

2.3 Games for Data Collection

Integrating insights from Games With a Purpose (GWAPs), News Ninja's design objectives aim to educate while simultaneously crowdsourcing linguistic bias labels, introducing a unexplored strategy in media bias research [23, 42]. Von Ahn [22] describes GWAPs as games that use human computation [43] to produce data during gameplay [22, 44, 45, 46, 47]. Lance et al. [48] view it as crowdsourcing via games. They offer inexpensive and scalable data collection [22] with an inclusive design that appeals to casual gamers [23]. Broadly, GWAPs are serious games, defined as games with an objective beyond pure entertainment. They often intersect with education or simulation [25] and can enhance educational scenarios [49, 50, 51, 52]. In contrast to gamification, which applies game elements in non-gaming contexts [53, 54], serious games are comprehensive gaming experiences. Subsequently, GWAPs can fall anywhere on the spectrum between serious games and gamification [55]. Thus, sustaining player engagement remains a challenge; GWAPs must be compelling, and players must be able to complete the task [56]. This stresses the need for engagement strategies to ensure sustainable, long-term data collection to account for changes in news content [23].

Historically, GWAPs that collect training data have been used for tasks associated with more objective truths, such as image labeling [57, 22] or grammatical parts-of-speech tagging [23, 58]. GWAPs that focus on subjective, cultural truths, such as detecting abusive, stereotypical, or sexist language, often merely describe their systems [59, 60] or conduct UX studies [61, 62]. Few evaluations manually assess the data [63, 62], use metrics like inter-annotator agreement (IAA), or compare with a domain-specific gold standard [64, 65], hindering direct comparison. Therefore, viable ways to evaluate a linguistic bias GWAP include assessing UX, manually evaluating the data, and comparing GWAP labels to gold standard datasets.

2.4 Combining Game Mechanics for Learning and Data Collection

Our goal is to merge learning with data collection within a single game, aiming for players to develop mental models for bias detection applicable across various tasks and in real-world scenarios [28]. We employ the "Gamified Knowledge Encoding Model" to leverage interacting game mechanics to help players internalize learning objectives [28]. Those game mechanics transform learning objectives into game elements while preserving the essence of fun and engagement [30]. Game mechanics serve as the interface between player interactions and learning outcomes, enclosing both game-bound elements tied to the storyline and game principles and player-bound actions executed by players. Through these interactions, players can acquire declarative knowledge and, with sufficient repetition, procedural knowledge [28].

To implement learning and data collection mechanisms, we analyze two games that already combine them [66, 67]. The language learning application Duolingo, originally designed to train users to translate language segments, faces similar challenges of training players to become experts

[19] and sustaining their engagement for continuous translation [67].[3] Duolingo resolves the conflict inherent in GWAPs - balancing data collection with providing challenging, educational, and entertaining experiences to players - by optimizing motivation through learning opportunities, gamification, challenges, and social nudges [23]. They further incorporate the pedagogical agent Duo [68], who guides players through the lessons. Pedagogical agents are essentially characters in a virtual learning landscape. They serve diverse instructional roles by providing help, guidance, and assistance in learning [69] through social cues that can cause social responses [70].

Similarly, the GWAP WordClicker educates players about word classes and gathers word class labels [66]. Players collect words of a chosen class to fill up jars used as resources for in-game currency production. They progressively learn to identify different classes by accumulating currency and buying new word classes. A time component increases the challenge for players. WordClicker serves as a tutorial for the more complex GWAP Tile Attack [56] and is part of Madge et al. [71]'s pipeline of games for part-of-speech tagging. Each game in the sequence augments complexity and cumulatively enhances player progression and engagement.

An effective linguistic bias GWAP integrates a structured, hands-on, expert-reviewed tutorial, complemented by learning mechanics with a cohesive feedback system and testing tasks for player qualification [42]. It involves setting explicit goals and weaving a compelling narrative, ideally situated within a context related to the learning objective [72]. The feedback enables iterative learning from errors and fosters improvement through repetition and correction, while testing tasks facilitate data selection for the final dataset. The learning material and tasks should be split into understandable units using didactic content structuring. For linguistic bias, we follow Spinde et al. [19] by taking sentences from news articles and collecting annotations at both the sentence and word levels. Players annotate sentences as "biased" or "not biased," while they can mark individual words as "biased" (Figure 3). To enhance player enjoyment, we incorporate game elements outlined for serious games and GWAPs by Segundo Díaz et al. [31], including feedback, progression elements such as levels, rewards like currency, and social interactions through discussion threads, as detailed in Section 3.6. The combination of a captivating interface and a progressively challenging GWAP, stressing the underlying purpose, can heighten player motivation and ensure their sustained engagement [73, 72, 29, 55].

Notably, the typical player base of GWAPs belongs to the *Achiever* or *Philanthropist* player group [74]. Implementing game mechanics that spotlight performance, milestones, progression, or the overarching mission drives their motivation [75, 29, 31]. Such mechanics can promote prolonged player engagement [29] by fostering enjoyable experiences and flow states [76]. Suited mechanics make the action steps harder, show progress and achievements, or unlock new content and interaction possibilities. While it is essential to design a linguistic bias GWAP with broad appeal to diversify the dataset, strategically fostering the motivations of *Achievers* and *Philanthropists* can amplify their contributions [55]. This becomes evident when considering that 3% of players are responsible for producing between 80-90% of the data [66]. Players' backgrounds impact bias perception, so GWAPs must include them to ensure a balanced, diverse, and minimally biased dataset [77]. The mission statement of the game, which explains the deeper purpose behind the annotation task and stresses the societal importance, can include the reasoning for querying player demographics. Stressing the learning value for players themselves can further increase intrinsic motivation.

[3]Initially, Duolingo's primary objective was translation. However, it pivoted to language learning with a subscription-based business model.

3 The News Ninja Game

News Ninja is a GWAP designed to educate players and collect linguistic bias annotations on words and sentences. The game translates written annotation guidelines into an interactive tutorial (Q1) and converts the annotation process into two data game mechanics (Q2). News Ninja's system design builds from four primary components: (1) Two data annotation mechanics (Section 3.1), (2) general game mechanics[4], and their integration within (3) the tutorial (Section 3.3) and (4) five game modes (Section 3.4). The two data mechanics extract annotations from player interactions to aggregate them into bias labels. The tutorial aims to increase players' bias detection skills, introducing the data mechanics to prepare players for the five game modes.

The design process of News Ninja adapts the Gamified Knowledge Encoding Model [28] in a four-step process, detailed in Figure 2. First, *knowledge* is divided into short units, each covering a single learning objective. Then, we fit game mechanics for the *moderation and mediation* of the units. The units' content is integrated through the pedagogical agent, demonstrating game modes, narrative, and repetition. In the next step, we design player-bound and game-bound *game mechanics* that allow for applying the new knowledge and frame the learning environment. The annotation mechanic is integrated as a player-bound mechanic. The interaction between game-bound and player-bound mechanics creates *learning affordances* and enables the formation of *mental models* through repeated interaction. Such *mental models* allow players to increase their game performance and detect bias in the real world.

Players start their journey, and the narrative introduces them as interns at a news outlet with an office plant as their pedagogical agent. The plant explains why media bias is an important issue players can help with and why their personal background matters. It then guides players through the demographic survey. Next, the plant leads them through the interactive tutorial (Section 3.3) with immediate feedback (Section 3.2) on the annotation mechanic (Section 3.1). Before the main gameplay, players encounter previously classified sentences, reinforcing learning through direct feedback and assessing their bias detection skill. Later, they unlock additional game modes and topics that incorporate social interactions and enable discussion.

3.1 Data Annotation Mechanic

News Ninja divides data annotations into sentence level and word level mechanics. This structure mirrors the structure of the BABE dataset [19], a commonly used [78] expert-level media bias dataset, which aims to cover linguistic bias at the lowest identifiable level and without the influence of article-level context. In BABE, sentences are labeled "biased" or "not biased." Each sentence can have biased words. Hence, News Ninja's first application focuses on linguistic bias on the sentence level and excludes the article level, collecting binary bias annotations on word and sentence levels. For sentence level annotation, pictured in Figure 3, a left swipe annotates a sentence as "not biased," while a right swipe annotates it as "biased." Alternatively, players can use designated buttons. For word level annotation, players select biased words by tapping on them. Mimicking BABE, players of News Ninja can mark words as "biased" and still annotate a sentence as "not biased."

3.2 Feedback

The game employs two types of feedback: direct feedback and delayed feedback (Figure 4). Direct feedback is activated when the ground truth of a sentence or word is known. Within this framework, "ground truth" refers to the label of a sentence or word. A sentence level label is attained either

[4]All game mechanics are detailed in Table 2 and explained in Section 3.6.

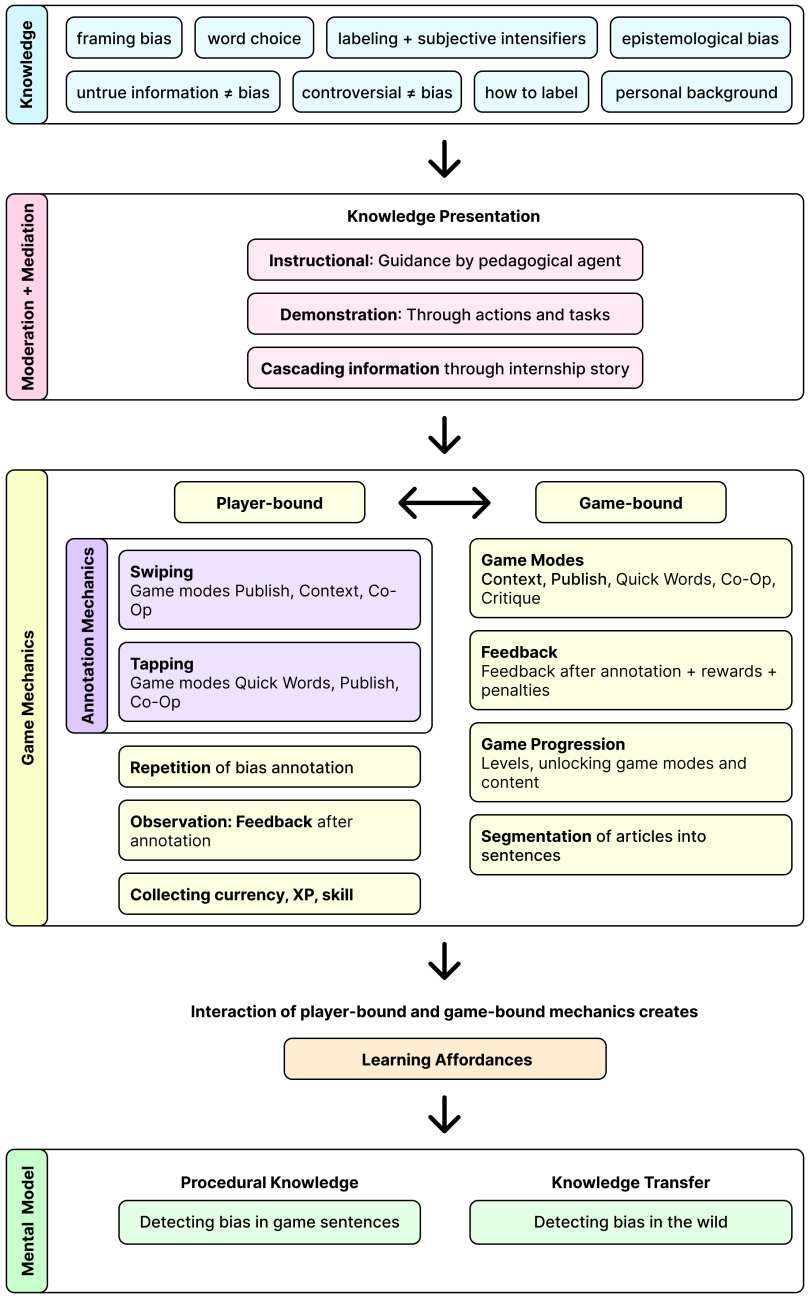

Fig. 2. The figure shows how News Ninja applies the Gamified Knowledge Encoding Model to gamify bias learning and facilitate annotation through four steps. First, News Ninja breaks down learning objectives into knowledge units. Those units are presented to players through game mechanics like the pedagogical agent, demonstrations, or the narrative. Next, the interaction between player-bound and game-bound mechanics creates learning affordances. These allow players, through repetition, to form mental models and apply their knowledge.

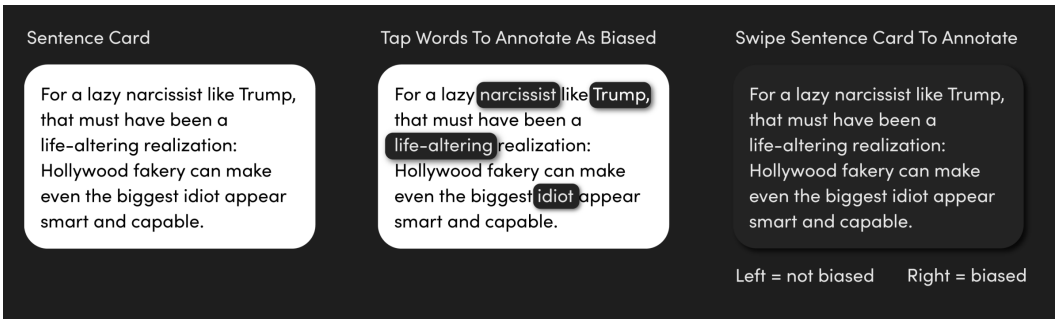

Fig. 3. Interaction and data annotation mechanic of the *Publish* game mode (Section 3.4.1). First, players tap biased words. Then, they swipe or use the buttons to annotate the sentence as "biased" or "not biased."

when the sentence originates from BABE [19], the baseline dataset,[5] or has at least five annotations by players who either voted "biased" or "not biased," thereby determining labels via majority vote [43].[6] Successful matches to the ground truth are rewarded with a green card outline, currency, skill, experience points, and sound effects; otherwise, the card turns red. The ground truth label, biased or not biased, appears above the card (Figure 1). For word level bias, tapped words are highlighted in green or red, corresponding to the ground truth match (Figure 4). The game considers a word "biased" when either two players (< 8 annotations) or 25% of players marked it as biased. Comparison against the ground truth and the hit/miss percentage calculation facilitates player rating, enabling the selection of inputs from players with higher bias detection rates.

However, for added sentences with no established ground truth, the "cold start problem" arises [58] as the game cannot give direct feedback. To navigate this challenge, News Ninja employs delayed feedback, visualized on the right in Figure 4. Here, players receive feedback consisting of yellow visual cues indicating they can revisit the game at a future point when sufficient data is available. The card outline turns yellow for delayed sentence level feedback, and a yellow dash icon appears. On the word level, selected word cards turn yellow. Then, the sentence moves to the *Paper* section. Players receive push notifications and see a yellow dot as a signifier on the navigation bar on the *Paper* section icon when ground truth is established, indicating new information and rewards. If players hit the ground truth, they receive a higher reward, while the uncertainty promotes extrinsic motivation [29]. Players increasingly encounter unclassified sentences with delayed feedback as they progress and increase their detection skills.

3.3 Tutorial

The tutorial progressively teaches linguistic bias through interactive examples and direct feedback while gradually increasing complexity at each level. Each tutorial level encapsulates one to two learning objectives without using scientific terms. Instead, the game aims for players to subconsciously learn to discern how bias manifests and identify it within sentence context. Players see ten manually selected sentences to classify in each tutorial level while receiving immediate feedback to foster learning. The tutorial starts with simple sentences and later transitions to more complicated ones. Similarly, the completion of each level unlocks a new game mode. The game modes incrementally increase the challenges by starting with the sentence level, progressing to the

[5]Although BABE has a lower agreement score relative to other datasets, it is notably high for a media bias dataset, a reflection of the inherent subjectivity in media bias that complicates achieving consensus, especially at the sentence and word levels. Moreover, it is the most comprehensive dataset currently available to us. We discuss this in Section 6.3.
[6]We further discuss mechanisms to determine labels in Section 6.4.

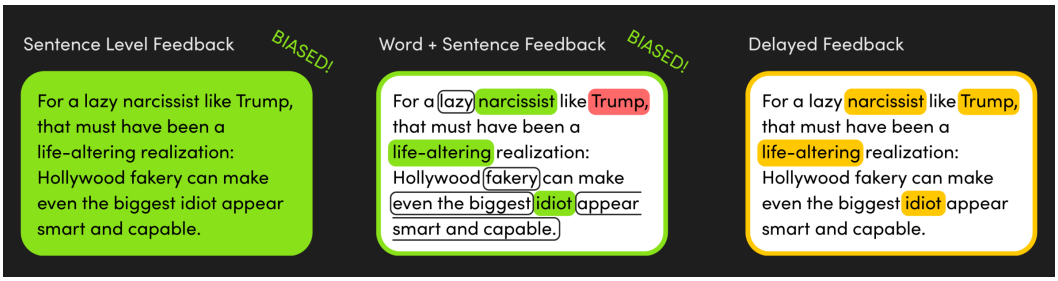

Fig. 4. Direct and delayed feedback in the *Publish* game mode (Section 3.4.1). Sentence level feedback color-codes the card. In this case, it turns green and indicates a hit. In case of a miss, it turns red. Next, correctly and incorrectly annotated words are shown in green and red. Missed words display a black outline. Stopwords do not count in the feedback and are displayed as right when surrounded by biased words. The third card shows delayed feedback. If a ground truth has not been established, sentence and word level feedback is displayed in yellow.

word level, and, ultimately, combining both. As a game element that is part of the UI, the plant visually grows and blossoms with the player's progress, symbolizing their learning.

We base the tutorial's content on the annotation guidelines for the expert-annotated dataset by Spinde et al. [19].[7] The tutorial mirrors these guidelines by covering the five elements of what media bias is, the importance of personal background, what needs to be annotated as biased, how to annotate, and what should not be annotated as biased.

The guidelines start with a general introduction to media bias, explaining that it can manifest through "particular word choice or framing exposing readers to non-neutral news reporting." For instance, the term "Coronavirus" is presented as unbiased, while "Chinese Virus" is considered biased. Further, it explains why it is important to be aware of linguistic bias and how it can manifest. The game's objective is simultaneously made clear: to train players to read with greater critical awareness and contribute to detecting bias.

Then, the guidelines explain that personal background impacts bias perception; hence, a demographic survey is necessary. News Ninja highlights the value of understanding players' viewpoints, followed by the demographic survey. It further emphasizes that players should set aside personal opinions on any topic, regardless of its political implications or alignment with their beliefs.

Next, the guidelines outline various types of bias:

- **Framing bias:** Skewing reader perception by only describing one point of view or frame.
- **Word choice:** Using one-sided terms or ideologically-driven depictions of concepts that alter readers' point of view.
- **Subjective intensiers**: Employing adjectives or adverbs that convey a strong opinion in that context, introducing bias.
- **Epistemological bias**: Manipulating language to affect the credibility of a statement, either enhancing or diminishing its believability.

News Ninja adapts this structure as the plant illustrates how framing can sway readers' opinions by presenting events from a single viewpoint. Subsequently, the plant discusses the impact of vague, dramatic, or sensational language and underscores how ambiguous or specific words can provoke emotional reactions. While the guidelines detail the annotation process, News Ninja opts for a more hands-on approach and demonstrates the game mechanics directly to players. Players are presented

[7]Derived on 20.07.23 from https://github.com/Media-Bias-Group/Neural-Media-Bias-Detection-Using-Distant-Supervision-With-BABE/blob/main/annotation_guidelines_BABE.pdf

with a sentence, tasked with identifying biased words, and receive direct feedback (Section 3.2). Next, the tutorial continues with epistemological bias.

Upon level advancement, a new game mechanic is introduced, focusing on the sentence's entire context and how a topic can be controversial without containing bias. The guidelines describe what should not be annotated as biased, stressing that controversial topics, or opinion-based reporting might not inherently be linguistically biased. For example, "abortion" is not a biased word, but the term "abortionist" is biased. The tutorial highlights that even statements containing untrue information do not necessarily feature biased language.

The final tutorial level explains the most complex annotation mechanic: Players first identify biased words by tapping and subsequently assess the entire sentence's bias using the swiping gesture or buttons. This level lets players practice the annotation mechanism while learning from direct feedback. Upon completion, it leads them back to the home screen.

3.4 Game Modes and UI

The game's home screen displays player statistics at the top, including in-game currency, experience points, and player level (Figure 1). A navigation bar on the bottom enables players to toggle between the home screen, the *Paper* section, a repository of sentences, both previously played and awaiting feedback, the community section, and the shop. The shop allows players to unlock new topics. Central to the home screen is the *Skill* bar and the *Mission* bar, explained in Section 3.6. Moreover, the *Breaking News* tile refreshes daily and showcases sentences in the *Publish* game mode (s. Section 3.4.1). Below are the tiles for the five game modes.

3.4.1 Game Modes. Effective bias detection relies on regular interaction with diverse content and feedback rather than pure theoretical understanding. The five game modes (1) *Context*, (2) *Publish*, (3) *Quick Words*, (4) *Co-Op*, and (5) *Critique* integrate the annotation mechanics differently to provide variety and cater to diverse player preferences. They also foster a sense of progression and achievement by unlocking new game modes. To increase fun, News Ninja introduces new elements into the mechanics, such as time constraints or cooperative challenges.

Post-tutorial, players can only access the (1) *Context* and (2) *Publish* game modes (Figure 1). Before playing, players select an available topic from which ten sentences are drawn.

The (1) *Context* mode operates with the sentence level mechanic. It shows a single sentence card to swipe with a $10 virtual currency reward for matching the ground truth. A left swipe annotates a sentence as "not biased," and a right swipe annotates it as "biased." After ten sentences, the game provides a summary, showing correct and incorrect classifications and permitting sentence review. News Ninja awards a bonus if players classify seven or more sentences correctly.

(2) *Publish* operates on sentence and word level by combining both annotation mechanics (Figure 1). Players first identify biased words by tapping on them before swiping the sentence card as they do in *Context*. Feedback includes missed biased words highlighted with a black border (4). We count correct words as a bonus and do not punish misses as it is often hard to find all biased words. Stopwords are automatically excluded from the calculation and shown as right if a biased word appears next to it.

The game assesses players before unlocking further game modes after the tutorial by presenting sentences with established ground truths and computing players' skill levels based on accuracy.

Next, the (3) *Quick Words* mode is unlocked. *Quick Words* focuses on word level annotation and adds a timed challenge (Figure 6). Players skip through sentences to tap as many biased words as possible before time runs out. Correct classifications earn game currency and additional time. Incorrect ones deduct time. If there is no majority vote yet, yellow feedback tiles show. A summary of identified words and their respective bias ratings shows when time runs out. *Quick Words*

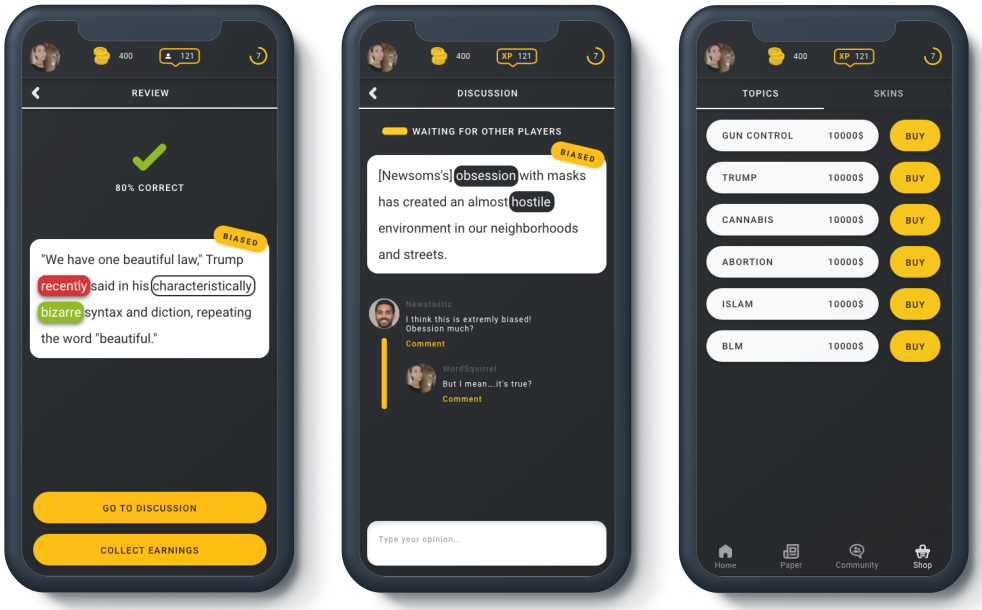

Fig. 5. The first screen (left to right) shows the detailed feedback for one correctly annotated sentence in the *Paper* section. It displays the sentence card with one word highlighted in red (incorrect annotation), one in green (correct annotation), and one with a black outline (missed word). Players can collect the reward or navigate to the discussion. The second screen shows the discussion of a sentence between two players. The sentence card shows on top with the comments below. The third screen shows the shop with six unlockable topics.

responds to research indicating greater word level than sentence level bias divergence among individual raters [19] and aims to increase word level annotations. While there is a higher risk of biased judgments when making quick, automatic decisions [79, 80], News Ninja prioritizes fun and player engagement to later monitor data quality more closely (s. Section 6.6).

The (4) *Co-Op* mode allows cooperative gameplay and integrates both annotation mechanics. Rewards are based on mutual agreement at both word and sentence classifications. The faster player receives a bonus. After the prior modes, we expect players to have achieved similar competencies, facilitating agreements.

The final (5) *Critique* mode includes both annotation mechanics by showing prior player annotations. Players can agree or disagree, adapt the ratings, and receive direct feedback when a ground truth forms (Figure 1). This game mode unlocks last because the game needs to ensure that players have collected enough experience to rate peers effectively.

3.4.2 Paper Section. Sentences from previous game rounds move to the *Paper* section (Figure 1). This section provides players with an opportunity to reflect on their prior gameplay. When a sentence with prior delayed feedback forms a ground truth, the game notifies players of the available feedback. Collecting this feedback — when in alignment with the ground truth — yields greater rewards than direct feedback, incentivizing players to revisit the game with an additional element of surprise.

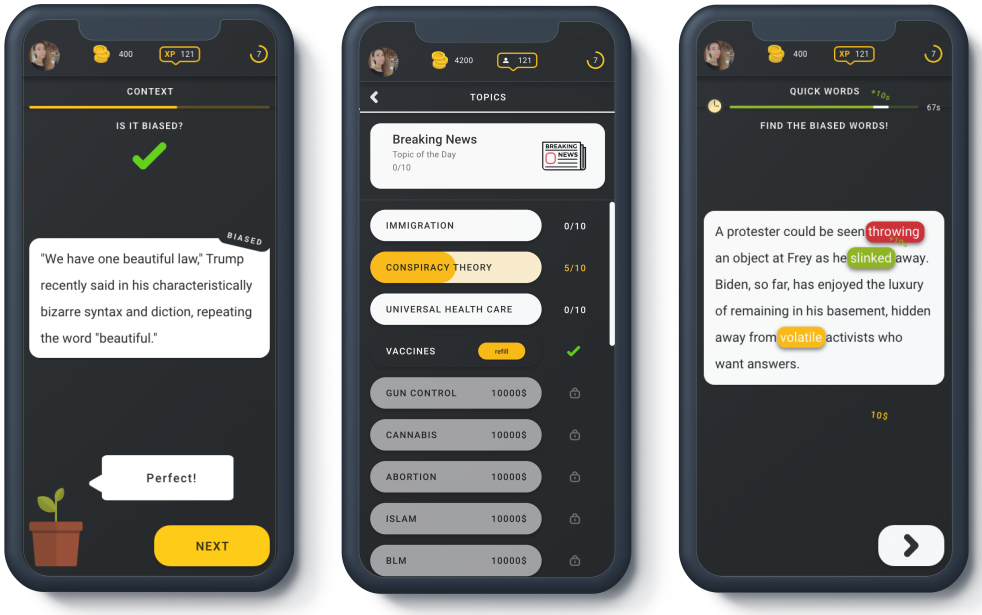

Fig. 6. The first left screen shows feedback in the game mode *Context*, with a sentence card displayed. It includes a green hook to indicate a correct answer on top, and the plant speaks motivationally in the left corner. The middle screen displays the topic selection before *Publish*. One topic was already played (black button with yellow refill button), and three are playable (white buttons). The right screen displays a *Quick Words* screen with the sentence card, feedback (green highlight for correct answer, red highlight for incorrect answer, yellow highlight for delayed feedback), the time bar, and the button for the next sentence in the bottom right corner.

3.5 Turning Player Input Into a Dataset

To turn player annotations into labels, News Ninja accumulates them from the game modes described in Section 3.1 on word and sentence level in the backend, as shown in Figure 7. Once a sentence or word reaches its thresholds, we assign a bias label based on a majority vote [43]. The resulting dataset contains sentence texts, sentence level labels, biased words within sentences, sentence topics, links to articles, the publishing outlet, and its respective leaning. The system enables training new bias classifiers with the game dataset using the approach of Spinde et al. [19]. In case of below-threshold annotations or a draw, players receive delayed feedback. A word receives a bias label if identified as such by a minimum of two players or by 25% of the players who encountered the sentence (Figure 7). Due to the challenges of identifying bias at the word level and the lower agreement reported in prior research [16], this threshold is deliberately set low. Annotations are only collected once a player surpasses the tutorial levels to ensure a basic understanding of linguistic bias. New sentences, including source and leaning in line with Spinde et al. [19], are added via a web application designed for content integration for continuous updates.

3.6 Motivational Game Elements

News Ninja employs various motivational game elements to sustain player engagement, detailed in Table 2. A sense of *progression* is achieved by buying new topics for sentences in the shop

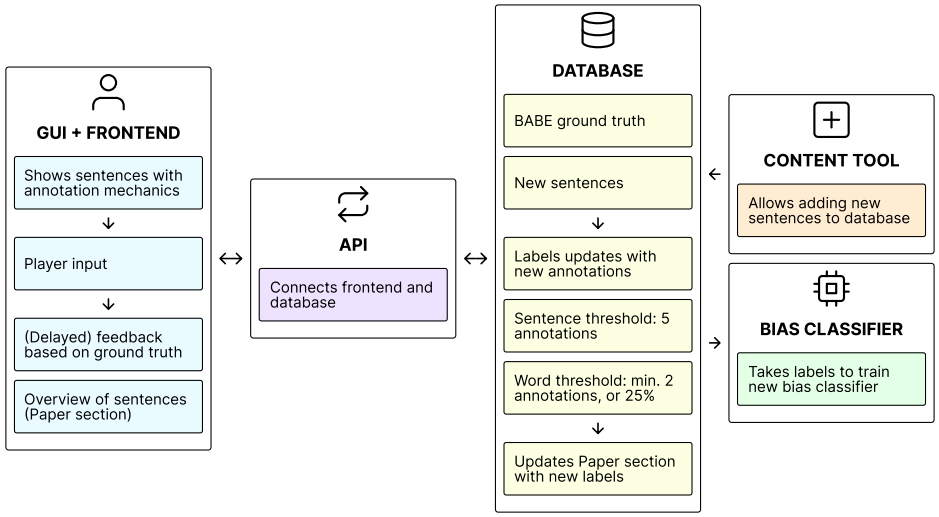

Fig. 7. System architecture of News Ninja. The game collects annotations from players in the front end, updates the labels in the backend, connects through the API, and shows feedback based on the new labels. The last step trains a new classifier with the obtained labels.

with in-game currency (Figure 5). By collecting experience points, players progress through levels and unlock new game modes [31]. Simultaneously, the plant grows with each level, symbolizing the players' growth. A progress bar on the home screen visualizes players' skill levels — players' accuracy on classified sentences — to increase intrinsic motivation by monitoring one's learning process. Beneath it, the group mission encourages a cooperative effort to mitigate bias by setting a number of labels as the goal that all players work on together, visualized through a progression bar below the skill level.

Daily, topics are refilled with limited sentences — capped at ten — to foster a feeling of *scarcity*, which makes topics more desirable [81]. The cappping also ensures players annotate all ten sentences and choose to annotate from different topics [29]. Players can purchase additional sentences using in-game currency (Figure 6). Similarly, the *Breaking News* tile on the home screen refills daily, offering higher rewards on daily completion (*Publish* game mode) and encouraging a build-up of consecutive play streaks. Further, it allows the developers to inject and prioritize sentences to balance the dataset selectively. Delayed feedback incentivizes players to return and collect rewards, reinforcing recurring player interaction through *unpredictability*. *Social motivation* is introduced through sentence discussions with other players (Figure 5).

3.7 Ethical considerations

The primary objective of the News Ninja game is to educate players about linguistic bias to collect bias labels through a gameful approach. While data collection is a significant aspect of this endeavor, we consciously avoid employing manipulative game design patterns [55, 82] aimed at coercing players to annotate more extensively. Instead, the game leverages players' intrinsic motivation to learn and contribute to mitigating bias, even at the expense of compromising the amount of data collected. Streak loss, the unpredictability of delayed feedback, and the capping of sentences are the only design patterns we consider more influencing in News Ninja.

4 Study Design

Initially, we conducted a qualitative pre-test of the tutorial (Q1) and the annotation mechanic (Q2) to assess players' experience (Section 4.1) and refine the game design (Section 3). We moved on to the primary study with crowdsourced workers to evaluate whether player-generated data achieves similar quality as expert labels (Q3, Section 4.2).

4.1 Pre-Test of Tutorial and Data Annotation Mechanic

To evaluate players' experience with the tutorial (Q1) and annotation mechanics (Q2), we conduct an A/B test with a UX survey with 21 participants on an early News Ninja mobile version. The game group played the game's tutorial, and the control group read through an introduction to biased wording used in prior studies [83] before annotating the same 20 sentences in the *Publish* game mode. The study aims to identify potential design flaws, assess whether the tutorial and game mechanics are enjoyable, understandable, and easy to use, and determine player motivation, likes, and dislikes. The UX survey incorporated the Single Ease Question (SEQ) [84] to query task difficulty, the 20-item Intrinsic Motivation Inventory (IMI) [85] to assess player motivation, and three open-ended questions regarding first impressions, general experience within the game, and any encountered problems (Section B). To evaluate differences in performance, we compare player annotations against the gold standard data set by Spinde et al. [19]. As our goal is to understand the experience of the game group more comprehensively, and since prior studies already used the control group's bias introduction, we randomly assigned 15 players to the game group and 6 to the control group.

We recruit volunteer participants through university group mail. Sixteen participants were between 20 and 29 years old, and five were between 30 and 40. One identified as women, 18 as men, and two as diverse. Regarding educational background, 14 held a bachelor's degree, five had completed graduate work, one had finished high school, and one had partially completed high school. Regarding language proficiency, 14 participants were fluent in English, and seven were at an intermediate level. The political orientation of the sample showed a left slant; 12 participants identified as leaning left, three as leaning right, and six as centrist. As for media consumption habits, 3 participants consumed news several times per day, six daily, nine several times per week, one several times per month, and two rarely.

The analysis revealed that the game group ($n = 15$) had a 8% greater alignment with the gold standard ($M_{\text{Game}} = .81, SD = .11; M_{\text{Control}} = .75, SD = .06$) compared to the control group ($n = 6$), suggesting the tutorial enhanced annotation quality. However, the increase is insignificant due to the small sample size. Feedback from the game group underscored a broad appreciation for the game's approach to media bias. Participants especially valued the guidance provided by the plant. The game group had a mean SEQ of 4.6/5 ($SD = .34$), and the control group had a mean of 3.6/5 ($SD = .87$). The IMI scored 5.2/7 ($SD = .54$) for the game group and 4.6/7 ($SD = .92$) for the control group. In line with the SEQ results, control group participants reported feeling overwhelmed by the task, indicating that the annotation guidelines from previous studies may have been too ambiguous. The open-question answers also highlighted a high complexity in the annotation mechanic. Other feedback expressed a wish to revisit prior annotations and pointed to the potential advantages of additional gamification and narrative integration. In response to the complexity noted by players, we refined the tutorial by simplifying the wording, shortening each lesson, and dividing the lessons into smaller sections followed by interactive examples. To simplify the game mechanics, we segmented them into five distinct game modes, described in Section 3.4, and added a section to review past and delayed sentences. Next, we develop the redesign as a mobile-first web application that does not require the installation of an app.

4.2 Study Outline

The final study takes participants through a streamlined game version, aiming to ensure comparability by minimizing potential other variables that could influence player behavior. The goal is to re-annotate and analyze the data quality of 10% of BABE sentences (370 sentences) and 150 new sentences, resulting in 520 sentences that each need at least five player annotations (2600+ annotations). To keep the study duration around 20 minutes, each player must play through the tutorial and 30 randomly and equally distributed sentences. Hence, the study requires 100 participants (100 Players ∗3 rounds ∗10 Sentences = 3000 Annotations) that we recruit from the US on the micro-tasking platform Prolific with a payment of 6£ per hour.

Participants began by reading the data processing agreement. On agreement, they continued to the game; otherwise, they were redirected to Prolific. Participants progress through the demographic survey[8] through which we assess players' backgrounds to monitor dataset bias. The survey includes questions on age, gender, nationality, education, political leaning, news consumption frequency, and English proficiency (Section A). Political orientation is important for assessing bias through slant, and English proficiency is essential for grasping linguistic nuances like bias.

The game has three phases. In phase 1 (tutorial), players learn about different types of linguistic bias and the game mechanics through the interactive tutorial described in Section 3.3. In phase 2 (direct feedback), players play 20 randomly drawn sentences in the *Publish* game mode (Section 3.4) and receive direct feedback. In phase 3 (delayed feedback), players saw ten new sentences with delayed feedback. Upon completion, participants are thanked for their contribution and guided back to Prolific for payment.

We use Krippendorff's α as the Inter-Annotator Agreement (IAA) metric in our initial data quality assessment. IAA measures the consensus among annotators on labeling tasks beyond what would be expected by chance alone. This metric is widely recognized for its reliability in evaluating dataset reliability [86] and is frequently used to analyze linguistic and media bias datasets [9]. We benchmark the IAA of News Ninja against the IAAs of two datasets within the domain of media bias. Firstly, we compare it to the crowdsourced dataset MBIC [35], generated by non-experts who received a textual introduction to media bias. This comparison is particularly relevant due to the similar recruitment methods through microtasking platforms. Secondly, we compare News Ninja's IAA with the currently most extensive, expert-curated dataset BABE, developed by students and researchers focusing on media bias [19].

Since IAA only measures agreement, we analyze the 370 re-labeled BABE sentences and the 150 new sentences by comparing them to newly created expert labels. Our objective is to assess the degree to which player labels align with expert labels, in extension determining the effectiveness of the tutorial. We further manually assess the types of sentences where players diverge from experts. In addition, we compare the original BABE labels with the new expert labels [19] to evaluate the suitability of BABE as a ground truth for player training.

4.3 Material

The BABE dataset [19] functions as both the evaluation benchmark and the ground truth for player training. To ensure comparability, participants re-annotate 10% of the original dataset (370 sentences). As relying solely on one dataset might add bias and the game's objective extends beyond mere re-labeling to create an extensive, crowdsourced linguistic bias dataset, two researchers compiled 150 new unlabeled sentences. Hence, we test if the system can generate new labels with sufficient quality. We limit the number of new sentences to 150 to keep the game duration around

[8]As we conduct the study on Prolific, participants agree to complete the entire survey. Future online versions will offer the option to skip each question.

20 minutes. All sentences, including those from the open-source dataset BABE, are sourced from publicly available data on news websites. The collection used the topics and timeframe of Spinde et al. [19] and leveraged AllSides[9] to ensure balanced political representation. Similar to Spinde et al. [19], the ratio of "biased" to "not biased" sentences is 2:1. Therefore, two datasets emerge from the study design: one with re-annotation sentences and one with new sentences.

4.4 Participants and Inclusion Criteria

The study involved 100 Prolific-recruited participants. We briefed prospective participants on the study's details, estimated duration, and compensation via Prolific. Those interested navigated to the React game application within their browsers (mobile or desktop). The platform's design ensured comprehensive data collection upon game completion. Thus, successful game completion served as the primary inclusion criterion. Additionally, participants self-reporting English proficiency below an intermediate level were excluded. We preliminary evaluated the study platform to test Prolific integration and data processing accuracy.

5 Results

5.1 Demographics

Participants had an average age of 36.72 years. Of all participants, 50% identified as men, 47% as women, and 3% as diverse. Every participant was of US nationality. 34% held a Bachelor's degree, 22% had some college education, 21% had completed high school, 12% had pursued graduate studies, 8% held an associate degree, 2% had undergone vocational or technical training, and 1% chose not to disclose their educational background. Political orientation was assessed with a 21-point scale (0 = left, 20 = right). Participants showed a left slant with an average score of 6.97. Most participants were frequent news consumers.[10] 31% consumed news daily, 29% multiple times a day, and 23% several times a week. Only 13% consumed news a few times a month, and 4% either never consumed news or did so very infrequently. Four participants reported intermediate English proficiency; all others indicated advanced proficiency. Thus, we exclude no participants based on proficiency. The average completion time was 21.45 minutes, with a median of 20 minutes.

5.2 IAA Assessment

The sentences extracted from the BABE dataset and annotated with the game interface achieve an IAA (Krippendorff's α) of 0.44. This IAA surpasses similar crowdsourced annotations, which reported $\alpha = 0.21$ [35] — an increase of 109.52%. Moreover, it outperforms the expert annotations that recorded a Krippendorff's α of 0.39 [19], an increase of 10.28%. Figure 8 shows a histogram of the bootstrapped game-annotated dataset (blue) and the expert dataset (orange). The 95% confidence intervals of the two datasets do not overlap, indicating a significant increase in Krippendorff's α between the two datasets.

The new set of sentences achieves Krippendorff's α of 0.399. This figure matches the IAA of expert annotations [19] and represents a 90% increase compared to crowdsourced annotations [35]. The significant increase in IAA indicates that News Ninja could generate labels on new sentences that were comparable to expert quality in terms of annotator agreement.

5.3 Comparison to Expert Labels

The player sentence labels achieve an accuracy of 79.8% with the new expert labels, with a precision of 95.5% and a recall of 69.2%. Players were more prone to missing actual positives but could reliably

[9] https://www.allsides.com/

[10] The types of news media, including digital, print, or TV, were not specified.

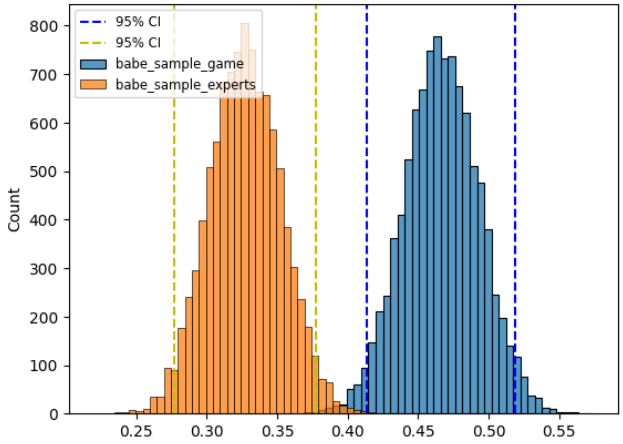

Fig. 8. Bootstrapped histogram of Krippendorff's α for game-annotated (blue) and expert-annotated (orange) datasets. The two confidence intervals do not overlap, indicating a significant increase in IAA.

Table 1. Confusion matrix of News Ninja labels and expert labels.

	Player Positive	**Player Negative**
Expert Positive	213	95
Expert Negative	10	202

avoid false positives (Table 1). From the 370 re-labeled BABE sentences, 289 player labels match the expert ratings, while 81 differ, resulting in an agreement rate of 78.11%. In analyzing the sentences with discrepancies, we find that 74 sentences (91.36%) identified as biased by experts were not labeled as biased by players. A manual analysis reveals that these sentences typically feature subtle linguistic bias, characterized by one or two bias-inducing words. For instance, the experts assigned a "biased" label to the following sentence, while players did not: "At the time, Wilson claimed he was only helping illuminate the contours of the 2016 race, before eventually being hired by one of the White House hopefuls." Here, the word "claimed" introduces bias by questioning the believability of Wilson [33].

Conversely, only seven sentences (8.64%) labeled unbiased by experts were labeled as biased by players. These sentences frequently involve controversial topics such as transgender issues, racial issues, abortion and pregnancy, immigration, or vaccines. In two cases, direct quotes contained strong words. While the quote is biased, the sentence reporting it is not biased. A comparison between player labels and BABE labels shows a 79.13% match, with 77 sentences receiving differing labels. The concordance between BABE and expert labels mirrors that of player and expert labels, standing at 61 sentences with an agreement of 83.47%. In some cases, player labels correct the labels from BABE. For instance, the sentence "And, of course, along the way, the hipster CEO is absolved of any responsibility to help uphold good cultural order." is labeled as "not biased" in BABE. However, both game labels and the new expert labels identify this sentence as biased.

Analyzing the 150 newly added sentences, we find an agreement of 84% with the expert standard. 126 sentences were labeled correctly, while 24 differed. Of the differing sentences, 21 were labeled as not biased (87.5%), while three were labeled as biased (12.5%).

6 Discussion

This paper presents News Ninja's design and first evaluation, a GWAP designed for linguistic bias education and data collection, potentially heightening players' bias awareness during future news reading. This first iteration tests data collection with annotation mechanics for crowdsourcing subjective truths, demonstrating that News Ninja achieves IAA levels and data quality comparable to expert datasets, suggesting its viability for creating linguistic bias datasets (Section 5.2 and Section 5.3). While isolating the effects of single game mechanics is challenging, we surmise that the interactive tutorial and annotation mechanics significantly increased the IAA through the immediate feedback. An extensive summary of News Ninja's game mechanics and their potential effects can be found in Table 2. The game-centric system is scalable and promising to be cost-effective by leveraging crowdsourced players over experts [4], It mitigates the risk of dataset obsolescence by periodically updating contents to capture changes in news, context, and perception over time.

6.1 Guidelines to Tutorial (Q1)

Distinct from prior research on media bias annotation, our study participants received instant feedback after written guidelines that fosters direct learning from their input. The tutorial possibly contributed to the higher IAA and agreement with the expert standard through the clear structure, storyline, pedagogical agent, and broken-down learning objectives. While we aim to keep the explanations as brief as possible, there is a risk that players may quickly skim or click through the text. Consequently, it is essential to test players after the tutorial and establish a baseline for their bias detection skills, which can be incorporated into models generating the bias labels. Even if players skip through the text, the feedback provided during the tutorial might contribute to their learning. However, as this study primarily focuses on evaluating the quality of player labels, learning effects must be examined separately (Section 6.6).

As the tutorial sentences are manually selected, they are straightforward examples that remain relevant over time. However, the game must undergo constant review to adapt to changes in language. We expect the topics and content to change faster than the language and expressions of bias. Concurrently, we must ensure their relevance by periodically reviewing the teaching content. Involving educational scientists in later game iterations will further support this aim.

6.2 Annotations as a Game Mechanic (Q2)

Preliminary testing revealed that some players found the initial *Publish* game mechanics complicated. Hence, the tutorial redesign introduces two game modes to prepare players. The increased IAA of the quantitative study at both word and sentence levels suggests its potential success. The exact impact of individual mechanics remains elusive, and isolated testing might be unproductive as they are integral to the game. Despite acknowledging the topic's importance, some players perceived the annotation task as work-like (Section 4.1). This raises concerns about motivation without financial incentives, necessitating a separate study to evaluate the player experience.

6.3 Data Quality (Q3)

In terms of IAA, the re-annotated dataset outperformed the expert dataset [4] by achieving a 10.28% higher IAA than BABE and a 109.52% higher IAA than the crowdsourced dataset. This speaks for the advantage of combining annotation mechanics with game-based learning, gamification, and a

feedback system [27, 72] over traditional data annotation methods like Excel tables or annotation tools, possibly resulting in high consistency among participants. We believe the game's approach makes understanding the task easier and more engaging than reading through annotation guidelines, which we will investigate in future work (Section 6.6).

As observed in Section 5.3, players trained with BABE sentences reproduced the BABE annotations with 79.13% agreement, indicating the training process's effectiveness. However, this also raises concerns regarding the necessity for well-balanced training data to ensure proper performance and prevent the reproduction of biases within the training material. We regard BABE as the ground truth, a notion that is inherently problematic given the subjective nature of bias and the questionable premise of a singular, definitive ground truth. BABE may fall short of achieving this quality, as seen in the discrepancy of 16.53% in expert ratings. The dataset mostly mislabels sentences as "not biased" that contain subtle linguistic bias. In addition to manually selecting tutorial sentences for future iterations, the sentences for all first six levels should undergo manual selection. Similarly, increasing the threshold for sentence label decisions could ensure that players receive precise feedback.

We consistently saw that players recognized sentences with high levels of bias as biased. Conversely, sentences with low levels of bias were often overlooked and not labeled as biased (Section 5.3). This issue may stem from the binary labeling system currently in place. When faced with uncertainty over whether content is biased, players might be inclined to categorize sentences as unbiased. Introducing a scale could reflect these nuances, although it requires adjustments to the game's mechanics and feedback system, particularly regarding rewards.

Moreover, News Ninja should closely monitor the labeling of sentences on topics commonly perceived as biased, including transgender rights, race, religion, women's rights, and queer issues. Previous NLP GWAP datasets contained stereotypes regarding gender and sexual orientation [87]. In News Ninja, experts could check sentences involving controversial topics or subtle biases, similar to the system described by Demartini, Mizzaro, and Spina [88]. One solution to further balance the dataset would be to oversample sentences with low bias content with the help of experts.

6.4 Cognitive Bias and Cultural Truth

News Ninja is the first GWAP to address the challenge of linguistic bias data creation. In this section, we present three possible influences on data quality: (1) Bias through agreement, (2) cognitive biases, and the implications of (3) cultural truths. The goal of building better classifiers for bias detection strongly influences News Ninja. It aims to enhance agreement (IAA) for classifier training, resulting in a conflict between increasing IAA and diversifying opinions (1). Hence, the design could increase convergence between player annotations, mainly through the tutorial and the feedback based on BABE's ground truth, possibly leading to dataset bias by reinforcing BABE annotation patterns.

Cognitive biases (2) can introduce potential irrationalities or deviations from normative decision-making processes [80] while annotating content for media bias. Especially during crowdsourcing tasks where the objective determination of "true" answers is often elusive [89], they can decrease the quality of the annotated data [90]. To address this problem, Draws et al. [77] propose a checklist to identify and mitigate cognitive biases in crowdsourcing tasks. Within News Ninja, we identified four specific biases that could potentially compromise the quality of our dataset:

(1) Self-interest bias, which may incline annotators to skew data in favor of their political beliefs or inattentively overlook subtle biases while prioritizing speed over accuracy.
(2) Groupthink bias can be reinforced by displaying majority votes to players, thus encouraging conformity.

(3) Availability bias might affect judgments through the preconception of stereotypes within sentences.
(4) The anchoring effect [80], potentially introduced via tutorial content that reflects BABE's ground truth, may also influence annotators' judgments.

Further, agreement on linguistic bias is a cultural truth (3), the consensus based on a group of people's beliefs — a judgment based on perception rather than an objective truth (e.g., if a word is a noun) [43]. We assume that there is a cultural agreement between people of the same group that can be identified by aggregation, despite individual deviations [91]. However, linguistic bias, political leaning, and cultural backgrounds separate players into different groups. Hence, a mere "ground truth" by a majority vote may be inadequate. Achieving 100% accuracy or agreement may be unrealistic, especially considering the evolving nature of language and meanings. A dynamically evolving dataset might capture these changes more accurately. Even experts might diverge, as seen in BABE's IAA, or bias the dataset, given their shared academic backgrounds. Therefore, we suggest educating diverse annotators about media bias to capture a societal average. Classifier results should be viewed as pointers, enabling analysis and augmenting the cognitive capabilities of news readers rather than replacing them.

We propose the following approaches to address the three challenges. First, we implemented a check through manual expert analysis and a comparative assessment of the annotated data (Section 5.3). Additionally, we integrated querying demographic metrics, such as political leanings, into the game [77]. News Ninja must ensure the inclusion of data from players exhibiting unique annotation patterns or perspectives [43, 35]. Hence, a more sophisticated latent class model could account for the noisy data caused by individual biases such as political ideology, cultural backgrounds, possible gender effects, task difficulty, or expertise [92, 93]. Incorporating these metrics into a probabilistic model and applying statistical hypothesis testing could identify systematic patterns that indicate the presence of cognitive biases [77].

Certain design decisions require revisions to better align with our objective of developing a dataset for model training that includes diverse perspectives. The current version is agreement-oriented. As discussed, striving solely for consensus may not effectively capture cultural truths. Consequently, we are transitioning to a process-oriented game design. This approach rewards player actions and sustained participation rather than consensus per se. *Co-op* enhance agreement between raters by rewarding the replication of existing annotation patterns. Instead of rewarding agreement, it could reward the annotation and provide players with a comparison afterward, potentially revealing the political leanings of their counterparts. Presently, skill levels are determined based on agreement with the established ground truth. Instead, News Ninja could incorporate specific test sentences with more objectively identifiable instances of bias for skill assessment purposes [77]. Additionally, experience, mainly measured by the number of annotations, could count toward skill level. Further, designing non-binary, more flexible annotation mechanics could increase diversification and accommodate the multifaceted nature of bias [9]. Subsequent analyses should determine the legitimacy of such contributions and develop approaches for their management.

6.5 Limitations

A significant limitation to consider is that monetary compensation was the predominant motivation for participation. This leads to uncertainties about the game's genuine appeal without financial incentives and is further addressed in Section 6.6. This study focused on data quality by comparing player and expert labels. However, we did not assess whether the tutorial and exercises improved players' bias detection skills. While we compared their labels to expert labels, we did not test their

bias detection skills before playing through the tutorial. Consequently, players' long-term and short-term learning outcomes must be evaluated in a separate study, as detailed in Section 6.6.

The research team's Western-centric bias and the dataset bias from Spinde et al. [4] may manifest in the News Ninja datasets. This potential influence could create a self-reinforcing loop, as players were given feedback from a dataset exhibiting similar biases [19]. Further, segmenting articles into statements and sentences is limiting, as they show up without the context of the whole article. The game, tested in a constrained and streamlined version, may only partially represent the full gameplay experience. As we focused on the data quality after the tutorial in this iteration, we did not employ other gameplay metrics or survey instruments.

While the participant pool demonstrated gender diversity, it lacked a balanced representation in age, education, and nationality, potentially skewing representation and elevating IAA scores. Moreover, the participants primarily came from Prolific, a platform recognized for its Western academic leanings. Hence, the demographics of the players who generated the annotations must be considered to avoid bias in the dataset.

Likewise, the preliminary UX study (Section 4.1), while offering insights into the complexity of the first game mechanic, has limitations. The increase in agreement with the expert standard among the game group is not significant, and the groups were unevenly distributed, with significantly more players in the game group. While focusing on the game group allowed for more qualitative insights, it limited the study's quantitative explanatory power. Nearly all participants identified as men and had a high level of education. Although the study identified and subsequently addressed some design flaws, the UX of News Ninja requires a more comprehensive analysis, as detailed in the following section.

6.6 Future Work

The following four next steps guide future research:

(1) The game offers potential as an educational tool, especially within academic settings. In our next step, we will study learning effects by conducting tests pre- and post-gameplay. We will further focus on extending the tutorial in cooperation with educational researchers and testing it with a student audience. This step includes manually re-selecting sentences for the first six levels. We focus on subtle linguistic bias in later lessons to resolve the issues with imprecise BABE labels and help players achieve expert-level bias detection. Further, we will enhance the narrative and rely on expert-vetted content instead of BABE labels.

(2) A comprehensive evaluation of player experience, fun, and motivation, particularly concerning the tutorial and each game mode. A UX study would pinpoint and correct UX discrepancies, refining the experience to mirror a genuine GWAP. To measure motivation and enjoyment, either a qualitative UX study with open-ended questions or a quantitative UX study, such as the Single Ease Question [84] to evaluate task difficulty or the Intrinsic Motivation Inventory [85] to measure player motivation, are well-suited [31].

(3) Launching an online version of the game to measure unpaid player interactions. This would provide insights into annotation efficiency and scalability without monetary incentives while creating an extensive dataset without further financial investment. In line with Madge et al. [94], we can measure player engagement by considering lifetime judgments, average judgments per player, average lifetime play, monthly active users, retention, and throughput. Incorporating pre- and post-tutorial assessments can also determine the tutorial's impact. The impact of delayed feedback on player retention is especially interesting. Longitudinal post-gameplay studies can shed light on lasting learning effects. We further need to evaluate

if quick, automatic decisions in the game modes *Co-Op* and *Quick Words* create stronger biases in the annotations [79].

(4) Applying a more sophisticated latent class model that integrates personal backgrounds, biases, experience, and skill levels [92, 93] to create labels. The game could, additionally to the player skill, select whose annotations to take into the final dataset based on player backgrounds to ensure a more diverse and balanced dataset, especially politically and culturally.

Future game iterations could benefit from integrating AI-driven explanations, such as OpenAI's ChatGPT.[11] Our experience shows that ChatGPT's ability to detect biased phrases is limited. However, it can provide explicit feedback by explaining why something may be considered biased after annotation tasks. While models such as GPT can generate labels, we believe assessing bias through human perception will be necessary, especially by including different opinions to develop and constantly evaluate a fair AI.

Like any online community, discussion threads in News Ninja require moderation. A report button and manual moderation of interactions are necessary, as models designed to detect hate speech might mistakenly label biased discussion content as hateful.

GWAPs, if executed proficiently, are a powerful tool for data-intensive, complex research areas if they succeed in making the process enjoyable. They may increase participation rates, improve data quality, and provide a valuable educational experience for players. Such methodologies can extend to various crowdsourced data collection tasks beyond linguistic bias detection. For example, we plan to use the News Ninja system to incorporate other types of media bias, misinformation, or manipulative language with corresponding lessons, storylines, and game modes.

7 Conclusion

This work introduces News Ninja, the first functional Game With A Purpose (GWAP) designed to educate players on detecting linguistic bias in news texts and to gather annotations to aid in automatic bias detection. News Ninja translates annotation guidelines into an interactive tutorial with direct feedback by applying frameworks from serious games and gamification. We describe the design and integration of the annotation task into different game mechanics and modes. Following a qualitative pre-test to assess player experience, a quantitative study was conducted to collect annotations via the game. The quality of annotations is evaluated by comparing player-generated labels against those from crowdsourcers and experts. The News Ninja dataset outperforms analogous linguistic bias datasets while achieving results comparable to experts, suggesting News Ninja as a promising approach for collecting annotations on linguistic bias. Furthermore, News Ninja exhibits potential for scalability, adaptability, and applications in educational settings.

Acknowledgments

This work was supported by the Hanns-Seidel Foundation (https://www.hss.de/), the German Academic Exchange Service (DAAD) (https://www.daad.de/de/), the Bavarian State Ministry for Digital Affairs in the project XR Hub (Grant A5-3822-2-16), and partially supported by JST CREST Grant JPMJCR20D3 Japan. None of the funders played any role in the study design or publication-related decisions.

[11] https://openai.com/blog/chatgpt

References

[1] Alexander Dallmann et al. "Media Bias in German Online Newspapers". In: *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. HT '15. New York, NY, USA: Association for Computing Machinery, 2015, pp. 133–137. ISBN: 9781450333955. DOI: 10.1145/2700171.2791057. URL: https://doi.org/10.1145/2700171.2791057.

[2] William P. Eveland Jr. and Dhavan V. Shah. "The impact of individual and interpersonal factors on perceived news media bias". In: *Political Psychology* 24.1 (2003), pp. 101–117. DOI: https://doi.org/10.1111/0162-895X.00318. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/0162-895X.00318.

[3] Pippa Norris. *A Virtuous Circle: Political Communications in Postindustrial Societies*. Communication, Society and Politics. Cambridge: Cambridge University Press, 2000. ISBN: 978-0-521-79015-4. DOI: 10.1017/CBO9780511609343. URL: https://www.cambridge.org/core/books/virtuous-circle/93623037EA261D4CA3AE0CB41E41A46A.

[4] Timo Spinde et al. "Automated identification of bias inducing words in news articles using linguistic and context-oriented features". In: *Information Processing & Management* 58.3 (Jan. 2021), p. 102505. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2021.102505. URL: https://www.sciencedirect.com/science/article/pii/S0306457321000157/pdfft?md5=64e81212b3bfa861d01a6fe3d5b979c3%5C&pid=1-s2.0-S0306457321000157-main.pdf.

[5] David Domke, Kelly McCoy, and Marcos Torres. "News media, racial perceptions, and political cognition". In: *Communication Research* 26.5 (1999), pp. 570–607. DOI: 10.1177/009365099026005003. URL: https://doi.org/10.1177/009365099026005003.

[6] Alberto Ardèvol-Abreu and Homero Gil de Zúñiga. "Effects of editorial media bias perception and media trust on the use of traditional, citizen, and social media news". In: *Journalism & Mass Communication Quarterly* 94.3 (2017), pp. 703–724. DOI: 10.1177/1077699016654684. URL: https://doi.org/10.1177/1077699016654684.

[7] Stefano DellaVigna and Ethan Kaplan. "The fox news effect: Media bias and voting". In: *The Quarterly Journal of Economics* 122.3 (Aug. 2007), pp. 1187–1234. ISSN: 0033-5533. DOI: 10.1162/qjec.122.3.1187. URL: https://doi.org/10.1162/qjec.122.3.1187.

[8] Jakob-Moritz Eberl, Hajo G. Boomgaarden, and Markus Wagner. "One bias fits all? Three types of media bias and their effects on party preferences". In: *Communication Research* 44.8 (2017), pp. 1125–1148. DOI: 10.1177/0093650215614364. URL: https://doi.org/10.1177/0093650215614364.

[9] Timo Spinde et al. "The Media Bias Taxonomy: A Systematic Literature Review on the Forms and Automated Detection of Media Bias". In: *arXiv* (2023). arXiv: 2312.16148 [cs.CL].

[10] Ceren Budak, Sharad Goel, and Justin M. Rao. "Fair and balanced? Quantifying media bias through crowdsourced content analysis". In: *Public Opinion Quarterly* 80.S1 (Apr. 2016), pp. 250–271. ISSN: 0033-362X. DOI: 10.1093/poq/nfw007. URL: https://doi.org/10.1093/poq/nfw007.

[11] Christoph Hube and Besnik Fetahu. "Neural based statement classification for biased language". In: *Proceedings of the twelfth ACM international conference on web search and data mining*. WSDM '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 195–203. ISBN: 978-1-4503-5940-5. DOI: 10.1145/3289600.3291018. URL: https://doi.org/10.1145/3289600.3291018.

[12] Sendhil Mullainathan and Andrei Shleifer. "Media Bias". In: *NBER Working Paper Series*. Cambridge, MA, Oct. 2002. DOI: 10.2139/ssrn.335800. URL: http://www.nber.org/papers/w9295.pdf.

[13] Timo Spinde et al. "Enabling News Consumers to View and Understand Biased News Coverage: A Study on the Perception and Visualization of Media Bias". In: *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. JCDL '20. Virtual Event, China: Association for Computing Machinery, Jan. 2020, pp. 389–392. ISBN: 978-1-4503-7585-6. DOI: 10.1145/3383583.3398619. URL: https://doi.org/10.1145/3383583.3398619.

[14] Filipe N. Ribeiro et al. "Media Bias Monitor : Quantifying Biases of Social Media News Outlets at Large-Scale". In: *Twelfth International AAAI Conference on Web and Social Media*. Palo Alto, California: AAAI Press, Jan. 2018, pp. 290–299. ISBN: 978-1-57735-798-8. URL: https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17878.

[15] Astrid Kause, Tarlise Townsend, and Wolfgang Gaissmaier. "Framing climate uncertainty: Frame choices reveal and influence climate change beliefs". In: *Weather, Climate, and Society* 11.1 (2019), pp. 199–215. DOI: 10.1175/WCAS-D-18-0002.1. URL: https://journals.ametsoc.org/view/journals/wcas/11/1/wcas-d-18-0002_1.xml.

[16] Timo Spinde et al. "How do we raise media bias awareness effectively? Effects of visualizations to communicate bias". In: *PLOS ONE* 17.4 (Jan. 2022). Publisher: Public Library of Science, pp. 1–14. DOI: 10.1371/journal.pone.0266204. URL: https://doi.org/10.1371/journal.pone.0266204.

[17] Timo Spinde, Felix Hamborg, and Bela Gipp. "An Integrated Approach to Detect Media Bias in German News Articles". In: *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. JCDL '20. Virtual Event, China: Association for Computing Machinery, Jan. 2020, pp. 505–506. ISBN: 978-1-4503-7585-6. DOI: 10.1145/3383583.3398585. URL: https://doi.org/10.1145/3383583.3398585.

[18] Yuanyuan Lei et al. "Sentence-level Media Bias Analysis Informed by Discourse Structures". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 10040–10050. URL: https://aclanthology.org/2022.emnlp-main.682.

[19] Timo Spinde et al. "Neural Media Bias Detection Using Distant Supervision With BABE - Bias Annotations By Experts". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Dominican Republic, Nov. 2021. DOI: 10.18653/v1/2021.findings-emnlp.101. URL: https://media-bias-research.org/wp-content/uploads/2022/01/Neural_Media_Bias_Detection_Using_Distant_Supervision_With_BABE___Bias_Annotations_By_Experts_MBG.pdf.

[20] Martin Wessel et al. "Introducing MBIB - the first media bias identification benchmark task and dataset collection". In: *Proceedings of 46th international ACM SIGIR conference on research and development in information retrieval (SIGIR23)*. New York, NY, USA: ACM, July 2023. DOI: https://doi.org/10.1145/3539618.3591882.

[21] Felix Hamborg, Karsten Donnay, and Bela Gipp. "Automated identification of media bias in news articles: an interdisciplinary literature review". In: *International Journal on Digital Libraries* 20.4 (Dec. 2019), pp. 391–415. ISSN: 1432-5012, 1432-1300. DOI: 10.1007/s00799-018-0261-y. URL: http://link.springer.com/10.1007/s00799-018-0261-y.

[22] Luis von Ahn. "Games with a Purpose". In: *Computer* 39.6 (June 2006), pp. 92–94. ISSN: 0018-9162. DOI: 10.1109/MC.2006.196. URL: http://ieeexplore.ieee.org/document/1642623/.

[23] Christopher Madge et al. "Incremental Game Mechanics Applied to Text Annotation". In: *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. Barcelona Spain: ACM, Oct. 2019, pp. 545–558. ISBN: 978-1-4503-6688-5. DOI: 10.1145/3311350.3347184. URL: https://dl.acm.org/doi/10.1145/3311350.3347184.

[24] Massimo Poesio et al. "Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation". In: *ACM Transactions on Interactive Intelligent Systems* 3.1 (Apr.

2013), pp. 1–44. ISSN: 2160-6455, 2160-6463. DOI: 10.1145/2448116.2448119. URL: https://dl.acm.org/doi/10.1145/2448116.2448119.

[25] Tarja Susi, Mikael Johannesson, and Per Backlund. *Serious Games : An Overview*. Institutionen för kommunikation och information, 2007. URL: https://urn.kb.se/resolve?urn=urn:nbn:se:his:diva-1279.

[26] Melisa Basol, Jon Roozenbeek, and Sander van der Linden. "Good News about Bad News: Gamified Inoculation Boosts Confidence and Cognitive Immunity Against Fake News". In: *Journal of cognition* 3.1 (Jan. 2020). Publisher: Ubiquity Press, pp. 2–2. ISSN: 2514-4820. DOI: 10.5334/joc.91. URL: https://pubmed.ncbi.nlm.nih.gov/31934684.

[27] Jeroen J. G. van Merriënboer Kirschner Paul A. *Ten Steps to Complex Learning: A Systematic Approach to Four-Component Instructional Design*. 3rd ed. New York: Routledge, Oct. 2017. ISBN: 978-1-315-11321-0. DOI: 10.4324/9781315113210.

[28] Sebastian Oberdorfer and Marc Erich Latoschik. "Gamified Knowledge Encoding: Knowledge Training Using Game Mechanics". In: *2018 10th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)*. ISSN: 2474-0489. Sept. 2018, pp. 1–2. DOI: 10.1109/VS-Games.2018.8493425.

[29] Yu-kai Chou. *Actionable Gamification: Beyond Points, Badges, and Leaderboards*. 2019th ed. Octalysis Media, Apr. 2015. ISBN: 978-1511744041.

[30] Sylvester Arnab et al. "Mapping learning and game mechanics for serious games analysis". In: *British Journal of Educational Technology* 46.2 (2015), pp. 391–411. ISSN: 1467-8535. DOI: 10.1111/bjet.12113. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/bjet.12113.

[31] Rosa Lilia Segundo Díaz et al. "Building blocks for creating enjoyable games—A systematic literature review". In: *International Journal of Human-Computer Studies* 159 (Mar. 2022), p. 102758. ISSN: 1071-5819. DOI: 10.1016/j.ijhcs.2021.102758. URL: https://www.sciencedirect.com/science/article/pii/S1071581921001762.

[32] Camiel J. Beukeboom and Christian Burgers. *Linguistic Bias*. July 2017. DOI: 10.1093/acrefore/9780190228613.013.439.

[33] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. "Linguistic Models for Analyzing and Detecting Biased Language". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 1650–1659. URL: https://aclanthology.org/P13-1162.

[34] Hannah Rashkin, Sameer Singh, and Yejin Choi. "Connotation Frames: A Data-Driven Investigation". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 311–321. DOI: 10.18653/v1/P16-1030.

[35] Timo Spinde et al. *MBIC – A Media Bias Annotation Dataset Including Annotator Characteristics*. arXiv:2105.11910. May 2021. DOI: 10.48550/arXiv.2105.11910. URL: http://arxiv.org/abs/2105.11910.

[36] Carl-Anton Werner Axelsson, Mona Guath, and Thomas Nygren. "Learning How to Separate Fake from Real News: Scalable Digital Tutorials Promoting Students' Civic Online Reasoning". In: *Future Internet* 13.3 (Mar. 2021), p. 60. DOI: 10.3390/fi13030060. URL: https://www.mdpi.com/1999-5903/13/3/60.

[37] John Cook, Stephan Lewandowsky, and Ullrich K. H. Ecker. "Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence". In: *PLOS ONE* 12.5 (May 2017). Publisher: Public Library of Science. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0175799. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0175799.

[38] Patricia Aufderheide. "Media Literacy: From a Report of the National Leadership Conference on Media Literacy". In: *Media Literacy in the Information Age* (2018).

[39] Sarah McGrew. "Learning to evaluate: An intervention in civic online reasoning". In: *Computers & Education* 145 (Feb. 2020), p. 103711. ISSN: 0360-1315. DOI: 10.1016/j.compedu.2019.103711. URL: https://www.sciencedirect.com/science/article/pii/S0360131519302647.

[40] David Buckingham. "Teaching media in a 'post-truth' age: fake news, media bias and the challenge for media/digital literacy education". In: *Cultura y Educación* 31 (May 2019), pp. 1–19. DOI: 10.1080/11356405.2019.1603814.

[41] Hendrik Heuer and Elena Leah Glassman. "A comparative evaluation of interventions against misinformation: Augmenting the WHO checklist". In: *Proceedings of the 2022 CHI conference on human factors in computing systems*. CHI '22. New York, NY, USA: Association for Computing Machinery, 2022. ISBN: 978-1-4503-9157-3. DOI: 10.1145/3491102.3517717. URL: https://doi.org/10.1145/3491102.3517717.

[42] Smilla Hinterreiter. "A Gamified Approach To Automatically Detect Biased Wording And Train Critical Reading". In: *2021 IEEE International Conference on Data Mining Workshops (ICDMW)*. Oct. 2021. DOI: 10.1109/ICDMW53433.2021.00141. URL: https://media-bias-research.org/wp-content/uploads/2021/10/hinterreiter2021a.pdf.

[43] Edith Law and Luis von Ahn. *Human computation*. ISSN: 1939-4616. Springer International Publishing, 2011. ISBN: 978-3-031-01555-7. DOI: 10.1007/978-3-031-01555-7. URL: http://dx.doi.org/10.1007/978-3-031-01555-7.

[44] Luis von Ahn and Laura Dabbish. "Designing games with a purpose". In: *Communications of the ACM* 51.8 (Aug. 2008), pp. 58–67. ISSN: 0001-0782, 1557-7317. DOI: 10.1145/1378704.1378719. URL: https://dl.acm.org/doi/10.1145/1378704.1378719.

[45] Luis von Ahn and Laura Dabbish. "Labeling images with a computer game". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '04. Vienna, Austria: Association for Computing Machinery, 2004, pp. 319–326. ISBN: 1581137028. DOI: 10.1145/985692.985733. URL: https://doi.org/10.1145/985692.985733.

[46] Edith L. M. Law et al. "TagATune: A Game for Music and Sound Annotation". In: *Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007, Vienna, Austria, September 23-27, 2007*. Ed. by Simon Dixon, David Bainbridge, and Rainer Typke. Austrian Computer Society, 2007, pp. 361–364. URL: http://ismir2007.ismir.net/proceedings/ISMIR2007%5C_p361%5C_law.pdf.

[47] Edith Law and Luis von Ahn. "Input-agreement: a new mechanism for collecting data using human computation games". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '09. Boston, MA, USA: Association for Computing Machinery, 2009, pp. 1197–1206. ISBN: 9781605582467. DOI: 10.1145/1518701.1518881. URL: https://doi.org/10.1145/1518701.1518881.

[48] Brent Lance et al. "Towards Serious Games for Improved BCI". In: Defense Technical Information Center, Jan. 2015, pp. 1–28. ISBN: 978-981-4560-52-8. DOI: 10.1007/978-981-4560-52-8_4-1.

[49] Frederik De Grove, Peter Mechant, and Jan Van Looy. "Uncharted Waters? Exploring Experts' Opinions on the Opportunities and Limitations of Serious Games for Foreign Language Learning". In: *Proceedings of the 3rd International Conference on Fun and Games*. Fun and Games '10. Leuven, Belgium: Association for Computing Machinery, 2010, pp. 107–115. ISBN: 9781605589077. DOI: 10.1145/1823818.1823830. URL: https://doi.org/10.1145/1823818.1823830.

[50] Richard Van Eck. "Digital Game Based Learning: It's Not Just the Digital Natives Who Are Restless". In: *EDUCAUSE* 41 (Jan. 2006).

[51] James Gee. "What Video Games Have to Teach Us About Learning and Literacy". In: *Computers in Entertainment* 1 (Oct. 2003), p. 20. DOI: 10.1145/950566.950595.

[52] Frank Greitzer, Olga Kuchar, and Kristy Huston. "Cognitive science implications for enhancing training effectiveness in a serious gaming context". In: *ACM Journal of Educational Resources in Computing* 7 (Nov. 2007). DOI: 10.1145/1281320.1281322.

[53] Sebastian Deterding et al. "From Game Design Elements to Gamefulness: Defining "Gamification"". In: *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*. MindTrek '11. Tampere, Finland: Association for Computing Machinery, 2011, pp. 9–15. ISBN: 9781450308168. DOI: 10.1145/2181037.2181040. URL: https://doi.org/10.1145/2181037.2181040.

[54] Jeanine Krath, Linda Schürmann, and Harald F. O. von Korflesch. "Revealing the theoretical basis of gamification: A systematic review and analysis of theory in research on gamification, serious games and game-based learning". In: *Computers in Human Behavior* 125 (Dec. 2021), p. 106963. ISSN: 0747-5632. DOI: 10.1016/j.chb.2021.106963. URL: https://www.sciencedirect.com/science/article/pii/S0747563221002867.

[55] Kathleen Tuite. "GWAPs: Games with a problem". In: *International Conference on Foundations of Digital Games*. 2014. URL: https://api.semanticscholar.org/CorpusID:10959426.

[56] Chris Madge et al. "Testing TileAttack with Three Key Audiences". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), May 2018. ISBN: 979-10-95546-10-8.

[57] M. Jordan Raddick et al. "Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers". In: *Astronomy Education Review* 9.1 (2010). Publisher: Portico. DOI: 10.3847/aer2009036. URL: https://doi.org/10.3847/aer2009036.

[58] Doruk Kicikoglu et al. "Wormingo: a 'true gamification' approach to anaphoric annotation". In: *Proceedings of the 14th International Conference on the Foundations of Digital Games*. San Luis Obispo California USA: ACM, Aug. 2019, pp. 1–7. ISBN: 978-1-4503-7217-6. DOI: 10.1145/3337722.3341868. URL: https://dl.acm.org/doi/10.1145/3337722.3341868.

[59] Federico Bonetti and Sara Tonelli. "A 3D Role-Playing Game for Abusive Language Annotation". In: *Workshop on Games and Natural Language Processing*. Ed. by Stephanie M. Lukin. Marseille, France: European Language Resources Association, May 2020, pp. 39–43. ISBN: 979-10-95546-40-5. URL: https://aclanthology.org/2020.gamnlp-1.6 (visited on 05/21/2024).

[60] Alice Millour et al. "Katana and Grand Guru: a Game of the Lost Words (DEMO)". In: *9th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC'19)*. Poznań, Poland, May 2019. URL: https://hal.science/hal-02106757 (visited on 05/21/2024).

[61] Fatima Althani, Chris Madge, and Massimo Poesio. "Less Text, More Visuals: Evaluating the Onboarding Phase in a GWAP for NLP". In: *Proceedings of the 9th Workshop on Games and Natural Language Processing within the 13th Language Resources and Evaluation Conference*. Ed. by Chris Madge. Marseille, France: European Language Resources Association, June 2022, pp. 17–27. URL: https://aclanthology.org/2022.games-1.3 (visited on 05/09/2024).

[62] François Bry and Marti Matthias. "Newsroom: A GWAP to Study Public Opinion". In: 2020. URL: https://api.semanticscholar.org/CorpusID:232216378.

[63] Ali Aghelmaleki. "Generating a non-sexist corpus through gamification for automatic sexism detection". PhD thesis. Mar. 2019. URL: https://kola.opus.hbz-nrw.de/frontdoor/index/index/year/2019/docId/1814 (visited on 05/21/2024).

[64] Daniele Vannella et al. "Validating and Extending Semantic Knowledge Bases using Video Games with a Purpose". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Kristina Toutanova and Hua Wu. Baltimore, Maryland: Association for Computational Linguistics, June 2014, pp. 1294–1304. DOI: 10.3115/v1/P14-1122. URL: https://aclanthology.org/P14-1122 (visited on 05/21/2024).

[65] Sarah Ita Levitan, Xinyue Tan, and Julia Hirschberg. "LieCatcher: Game Framework for Collecting Human Judgments of Deceptive Speech". In: *Proceedings of the 2020 International Conference on Multimodal Interaction*. Virtual Event Netherlands: ACM, Oct. 2020, pp. 762–763. ISBN: 978-1-4503-7581-8. DOI: 10.1145/3382507.3421166. URL: https://dl.acm.org/doi/10.1145/3382507.3421166 (visited on 05/09/2024).

[66] Chris Madge et al. "Making Text Annotation Fun with a Clicker Game". In: *Proceedings of the 14th International Conference on the Foundations of Digital Games*. FDG '19. New York, NY, USA: Association for Computing Machinery, 2019. ISBN: 978-1-4503-7217-6. DOI: 10.1145/3337722.3341869. URL: https://doi.org/10.1145/3337722.3341869.

[67] Luis von Ahn. "Duolingo: learn a language for free while helping to translate the web". In: *Proceedings of the 2013 international conference on Intelligent user interfaces*. IUI '13. New York, NY, USA: Association for Computing Machinery, Mar. 2013, pp. 1–2. ISBN: 978-1-4503-1965-2. DOI: 10.1145/2449396.2449398. URL: https://doi.org/10.1145/2449396.2449398.

[68] Amy L. Baylor and Yanghee Kim. "Pedagogical Agent Design: The Impact of Agent Realism, Gender, Ethnicity, and Instructional Role". In: *Intelligent Tutoring Systems*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2004, pp. 592–603. ISBN: 978-3-540-30139-4. DOI: 10.1007/978-3-540-30139-4_56.

[69] Ati Suci Dian Martha and Harry B. Santoso. "The Design and Impact of the Pedagogical Agent: A Systematic Literature Review". In: *Journal of Educators Online* 16.1 (Jan. 2019). DOI: 10.9743/jeo.2019.16.1.8. URL: https://eric.ed.gov/?id=EJ1204376.

[70] Robert K. Atkinson, Richard E. Mayer, and Mary Margaret Merrill. "Fostering social agency in multimedia learning: Examining the impact of an animated agent's voice". In: *Contemporary Educational Psychology* 30.1 (2005). Publisher: Elsevier Science, pp. 117–139. ISSN: 1090-2384. DOI: 10.1016/j.cedpsych.2004.07.001.

[71] Chris Madge et al. "LingoTowns: A Virtual World For Natural Language Annotation and Language Learning". In: *Extended Abstracts of the 2022 Annual Symposium on Computer-Human Interaction in Play*. CHI PLAY '22. New York, NY, USA: Association for Computing Machinery, Nov. 2022, pp. 57–62. ISBN: 978-1-4503-9211-2. DOI: 10.1145/3505270.3558323. URL: https://doi.org/10.1145/3505270.3558323.

[72] Sebastian Oberdörfer. "Better Learning with Gaming: Knowledge Encoding and Knowledge Learning Using Gamification". PhD thesis. Jan. 2021, p. 198. DOI: 10.25972/OPUS-21970. URL: https://www.researchgate.net/publication/349140377_Better_Learning_with_Gaming_Knowledge_Encoding_and_Knowledge_Learning_Using_Gamification.

[73] Raph Koster. *Theory of Fun for Game Design*. 2nd. O'Reilly Media, Inc., 2013. ISBN: 1-4493-6321-0.

[74] Gustavo F. Tondello et al. *The Gamification User Types Hexad Scale*. CHI PLAY '16. Austin, Texas, USA: Association for Computing Machinery, 2016, pp. 229–243. ISBN: 9781450344562. DOI: 10.1145/2967934.2968082. URL: https://doi.org/10.1145/2967934.2968082.

[75] Gustavo F. Tondello, Alberto Mora, and Lennart E. Nacke. "Elements of Gameful Design Emerging from User Preferences". In: *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. CHI PLAY '17. New York, NY, USA: Association for Computing Machinery, Oct. 2017, pp. 129–142. ISBN: 978-1-4503-4898-0. DOI: 10.1145/3116595.3116627. URL: https://doi.org/10.1145/3116595.3116627.

[76] Mihaly Csikszentmihalyi. *Flow: The Psychology of Optimal Experience*. Harper & Row, Jan. 1990.

[77] Tim Draws et al. "A Checklist to Combat Cognitive Biases in Crowdsourcing". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 9 (Oct. 2021), pp. 48–59.

ISSN: 2769-1349. DOI: 10.1609/hcomp.v9i1.18939. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/18939.

[78] Jakub Piskorski et al. *News Categorization, Framing and Persuasion Techniques: Annotation Guidelines*. Tech. rep. JRC-132862. Ispra (Italy): European Commission Joint Research Centre, Mar. 2023. URL: https://knowledge4policy.ec.europa.eu/text-mining/news-categorization-framing-persuasion-techniques-annotation-guidelines_en.

[79] Patricia L. Moravec, Antino Kim, and Alan R. Dennis. "Appealing to sense and sensibility: System 1 and system 2 interventions for fake news on social media". In: *Information Systems Research* 31.3 (2020), pp. 987–1006. DOI: 10.1287/isre.2020.0927. URL: https://doi.org/10.1287/isre.2020.0927.

[80] Amos Tversky and Daniel Kahneman. "Judgment under uncertainty: Heuristics and biases". In: *Science (New York, N.Y.)* 185.4157 (1974), pp. 1124–1131. DOI: 10.1126/science.185.4157.1124. URL: https://www.science.org/doi/abs/10.1126/science.185.4157.1124.

[81] Juho Hamari. *Gamification - Motivations & Effects*. ISSN: 1799-4942 (electronic). Aalto University, 2015. ISBN: 978-952-60-6056-9. URL: https://aaltodoc.aalto.fi/handle/123456789/15037 (visited on 05/07/2024).

[82] Colin M. Gray et al. "The Dark (Patterns) Side of UX Design". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Apr. 2018, pp. 1–14. ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3174108. URL: https://dl.acm.org/doi/10.1145/3173574.3174108.

[83] Timo Spinde et al. "TASSY - A Text Annotation Survey System". In: *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. Sept. 2021. DOI: 10.1109/JCDL52503.2021.00052. URL: https://media-bias-research.org/wp-content/uploads/2022/01/Spinde2021c.pdf.

[84] Jeff Sauro and Joseph S. Dumas. "Comparison of three one-question, post-task usability questionnaires". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '09. New York, NY, USA: Association for Computing Machinery, Apr. 2009, pp. 1599–1608. ISBN: 978-1-60558-246-7. DOI: 10.1145/1518701.1518946. URL: https://doi.org/10.1145/1518701.1518946.

[85] E. McAuley, T. Duncan, and V. V. Tammen. "Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: a confirmatory factor analysis". In: *Research Quarterly for Exercise and Sport* 60.1 (Mar. 1989), pp. 48–58. ISSN: 0270-1367. DOI: 10.1080/02701367.1989.10607413.

[86] Andrew F. Hayes and Klaus Krippendorff. "Answering the Call for a Standard Reliability Measure for Coding Data". In: *Communication Methods and Measures* 1.1 (Apr. 2007), pp. 77–89. ISSN: 1931-2458. DOI: 10.1080/19312450709336664. URL: https://doi.org/10.1080/19312450709336664.

[87] Jahna Otterbacher. "Crowdsourcing Stereotypes: Linguistic Bias in Metadata Generated via GWAP". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15. New York, NY, USA: Association for Computing Machinery, Apr. 2015, pp. 1955–1964. ISBN: 978-1-4503-3145-6. DOI: 10.1145/2702123.2702151. URL: https://doi.org/10.1145/2702123.2702151.

[88] Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. "Human-in-the-loop artificial intelligence for fighting online misinformation: Challenges and opportunities". In: *The Bulletin of the Technical Committee on Data Engineering* 43 (Sept. 2020), pp. 65–74.

[89] Lora Aroyo et al. "Crowdsourcing Subjective Tasks: The Case Study of Understanding Toxicity in Online Discussions". In: *Companion Proceedings of The 2019 World Wide Web Conference*. WWW '19. New York, NY, USA: Association for Computing Machinery, May

2019, pp. 1100–1105. ISBN: 978-1-4503-6675-5. DOI: 10.1145/3308560.3317083. URL: https://doi.org/10.1145/3308560.3317083.

[90] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. "Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. New York, NY, USA: Association for Computing Machinery, May 2019, pp. 1–12. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300637. URL: https://doi.org/10.1145/3290605.3300637.

[91] A. Kimball Romney, Susan C. Weller, and William H. Batchelder. "Culture as Consensus: A Theory of Culture and Informant Accuracy". In: *American Anthropologist* 88.2 (1986). Publisher: American Anthropological Association, Wiley, pp. 313–338. ISSN: 0002-7294. URL: https://www.jstor.org/stable/677564.

[92] Peter Welinder et al. "The multidimensional wisdom of crowds". In: *Advances in neural information processing systems*. Vol. 23. Curran Associates, Inc., 2010. URL: https://proceedings.neurips.cc/paper_files/paper/2010/file/0f9cafd014db7a619ddb4276af0d692c-Paper.pdf.

[93] Yan Yan et al. "Active learning from crowds". In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. ICML'11. Madison, WI, USA: Omnipress, June 2011, pp. 1161–1168. ISBN: 978-1-4503-0619-5.

[94] Chris Madge et al. "Experiment-Driven Development of a GWAP for Marking Segments in Text". In: *Extended Abstracts Publication of the Annual Symposium on Computer-Human Interaction in Play*. CHI PLAY '17 Extended Abstracts. New York, NY, USA: Association for Computing Machinery, 2017, pp. 397–404. ISBN: 9781450351119. DOI: 10.1145/3130859.3131332. URL: https://doi.org/10.1145/3130859.3131332.

A Demographic Survey

(1) What gender do you identify with? (Woman, Man, Diverse, Prefer not to say)
(2) What is your age? (Input field for number)
(3) What is the highest level of education you have completed? (8th grade, Some high school, High school graduate, Vocational or technical school, Some college, Associate degree, Bachelor's degree, Graduate work, Ph.D., I prefer not to say)
(4) What is the level of your English proficiency? (Proficient, Independent, Basic)
(5) Do you consider yourself to be liberal, conservative, or somewhere in between? Please slide to record your response. (Very liberal to Very conservative, -10 to 10 point slider)
(6) How often on average do you check the news? (Never, Very rarely, Several times per month, Several times per week, Every day, Several times per day)
(7) What news outlets do you consume? (Selection through checkboxes with free text field below)

B UX Study Open Questions

(1) What was your first impression when you entered the game?
(2) How was your experience within the game?
(3) Where did you struggle?

C Game Mechanics

Table 2. Game mechanics employed by News Ninja and their purpose and potential effects, with higher engagement translating to higher amounts of collected data.

Game Mechanic	**News Ninja Adaptation**	**Purpose and Possible Effects**
Direct Feedback	Color-codes outlines and highlights on word and sentence level	Enable players to learn from their input and repetition; Increase bias detection skills and IAA
Delayed feedback and uncertainty	Yellow outlines and highlights on word and sentence level; Notification and higher reward on formation of ground truth	Inform players that ground truth hasn't formed yet; Increase motivation to return
Guidance through pedagogical agent	Plant guiding through the tutorial and speaking motivating to players during gameplay	Transmit learning objectives; Increase bias detection skills, IAA, and motivation
Narrative	Plant tells story of player as intern in a news outlet	Increase enjoyment, motivation and learning effects through context related to the learning objectives; Increase understanding of media bias and engagement
Tutorial	Slowly increasing complexity through levels and new mechanics	Sense of progression and achievement; Increase bias detection skills and IAA
Appeal to higher meaning	Stating the purpose of News Ninja and effect in the world	Increase intrinsic altruistic motivation and engagement
Rewards and penalties	Experience points, in-game currency, rise in skill level, time penalties	Increase fun, extrinsic motivation, learning effects, and engagement
Ownership: Assessment	Skill level bar, feedback after game round, feedback after annotation	Show players their capabilities and increase of them over time; Increase intrinsic motivation and engagement
Progression	Level, Skill level bar, unlocking content and game modes	Increase fun and intrinsic and extrinsic motivation; Increase engagement and bias detection skills
Collaboration and Competition	Game mode *Co-Op*; Group Mission	Increase fun and engagement through social collaboration

Table 2. (continued)

Game Mechanic	**News Ninja Adaptation**	**Purpose and Possible Effects**
Time pressure	Game mode *Quick Words*	Increase fun and annotation collection on word level
Ownership: Collecting	Experience points, in-game currency, rise in skill level	Increase fun, extrinsic motivation, and engagement
Responsibility	Stating the game's mission, especially through the group mission	Increase intrinsic motivation and engagement
Discussion	Discussing bias annotations of sentences with other players	Reflection and motivation through social interactions; Increase fun and intrinsic motivation, potentially increasing bias detection skills

Received 21 February 2024; revised 03 June 2024; accepted 5 July 2024