Behaviour & Information Technology

ISSN: 0144-929X (Print) 1362-3001 (Online) Journal homepage: www.tandfonline.com/journals/tbit20

# Quest for quality: a review of design knowledge on gamified AI training data annotation systems

Simon Krohmann, Sascha Rank, Manuel Schmidt-Kraepelin, Scott Thiebes & Ali Sunyaev

**To cite this article:** Simon Krohmann, Sascha Rank, Manuel Schmidt-Kraepelin, Scott Thiebes & Ali Sunyaev (25 Oct 2025): Quest for quality: a review of design knowledge on gamified AI training data annotation systems, Behaviour & Information Technology, DOI: 10.1080/0144929X.2025.2568932

**To link to this article:** https://doi.org/10.1080/0144929X.2025.2568932



Published online: 25 Oct 2025.

BEHAVIOUR & INFORMATION TECHNOLOGY
https://doi.org/10.1080/0144929X.2025.2568932



# Quest for quality: a review of design knowledge on gamified AI training data annotation systems

Simon Krohmann[a,b], Sascha Rank[a], Manuel Schmidt-Kraepelin[a], Scott Thiebes[c] and Ali Sunyaev[d]

[a]Karlsruhe Institute of Technology, Institute of Applied Informatics and Formal Description Methods, Karlsruhe, Germany; [b]HIDSS4Health – Helmholtz Information and Data Science School for Health, Heidelberg, Germany; [c]School of Economics and Management, Advanced Institute of Business, Tongji University, Shanghai, China; [d]TUM School of Computation, Information and Technology, Technical University Munich, Heilbronn, Germany

**ABSTRACT**

Data annotation is a tedious, yet vital task to create AI training data. Gamification can support annotation quality by motivating and engaging annotators. Existing gamified annotation system artifacts are largely instantiations; it remains unclear how the ingrained design knowledge transcends individual and situational factors and can inform the design of gamified annotation systems as a system class. This study synthesises extant gamified annotation system artifacts and gauges the maturity of ingrained design knowledge. We conduct a semi-systematic review of 56 articles which present gamified annotation system artifacts. Beyond a broad overview of design artifacts and research activities, we derive 13 solution streams that describe design knowledge as means-end relationships between goals and gamification-based solutions. While some solution streams exhibit high maturity, the results largely confirm our initial assumption that the design knowledge base on gamified annotation systems is immature. To advance maturity, we recommend to move beyond creating new instantiations as the primary research activity and to focus on developing actionable design prescriptions. Our study contributes to creating more mature design knowledge on gamified annotation systems and offers valuable perspectives on the maturity and scholarly progression of gamification research.

**ARTICLE HISTORY**
Received 20 November 2024
Accepted 13 September 2025

**KEYWORDS**
Gamification; AI training data; annotation; design knowledge; literature review

## 1. Introduction

Consider an autonomous vehicle. For it to drive safely, the underlying artificial intelligence (AI) model must be trained with many street images in which human annotators have meticulously marked objects of interest such as other cars, street signs, or pedestrians. This process of adding metadata to existing data instances is called data annotation (Pustejovsky and Stubbs 2012) and is crucial for the training of AI models (Gudivada, Apon, and Ding 2017). For human annotators, data annotation tasks can be time-intensive, repetitive, and tedious; thus they do not easily afford sustained motivation and engagement (Warsinsky et al. 2022). However, if annotators are not motivated or engaged in the annotation task, issues like careless mistakes or sloppy annotations arise, which may ultimately reduce annotation quality (Chandler and Kapelner 2013; Neves and Ševa 2021). To ensure high annotation quality, effective ways to motivate and engage annotators are required. One purposive design strategy to do so is gamification, describing the use of game design elements like points, badges or leaderboards to evoke gameful experiences and thereby motivate individuals to perform certain tasks or behaviours (Koivisto and Hamari 2019).

Research on gamified annotation systems features a wide variety of gamification designs. While these offer valuable insights into how a gamification design for annotation tasks may look, feel, and behave, individual and situational factors vary greatly across research: studies feature different gamification elements (e.g. points; Mekler et al. (2017), or narratives; Dumitrache et al. (2013)), annotation purposes (e.g. improve disease classification; Balducci and Buono (2018), or cultural heritage preservation; Ivanjko (2019)), annotators (e.g. experts; Warsinsky et al. (2022), or a crowd; Feyisetan et al. (2015)), and created annotations (e.g. segmentation masks; Jauer, Spicher, and Deserno (2021), or tags; Mekler et al. (2017)). Gamification research has repeatedly stressed that the effects of gamification are highly sensitive to such contextual factors (Koivisto

**CONTACT** Simon Krohmann simon.krohmann@kit.edu Karlsruhe Institute of Technology, Institute of Applied Informatics and Formal Description Methods, Karlsruhe, 76133, Germany; HIDSS4Health – Helmholtz Information and Data Science School for Health, Heidelberg, Germany

and Hamari 2019; Nacke and Deterding 2017). Yet, the presented gamification design artifacts in literature are largely instantiations (i.e. real-world working systems; Hevner et al. (2004)), which are strongly tied to individual and situational factors (Gregor and Hevner 2013). More abstract design artifacts such as models or design principles (Gregor and Hevner 2013) that transcend individual and situational factors are scarce. Some studies go beyond instantiations and present frameworks (Friðriksdóttir and Einarsson 2022; L'Heureux et al. 2017) or architectures (Dumitrache et al. 2013) to gamify annotation tasks. However, the scope of these artifacts is limited to, for example, specific annotation tasks (e.g. text annotation; Dumitrache et al. 2013) or crowdsourcing approaches (L'Heureux et al. 2017). Moreover, while there exist syntheses of extant design knowledge in related domains like gamified crowdsourcing (Morschheuser et al. 2017), gamified learning technologies (Carmichael, MacEachen, and Archibald 2022; Ertan and Kocadere 2022), or (non-gamified) annotation tools (Sager, Janiesch, and Zschech 2021), we are not aware of any existing syntheses of design knowledge for gamified annotation systems as a system class. Given the lack of mature design artifacts that transcend individual and situational factors, and the lack of syntheses of extant design knowledge, it is difficult to extract abstract design knowledge from research on gamified annotation systems. Therefore, the principal aim of this study is to synthesise current gamified annotation system artifacts, to comprehensively structure the ingrained design knowledge, and to assess its maturity. By doing so, we seek to pave the way to derive more complete and mature design knowledge on gamified annotation systems. We raise the research question: *How mature is design knowledge on gamified AI training data annotation systems?*

To answer our research question, we conduct a semi-systematic literature review of 56 scholarly articles. Based on a systematic database search, we synthesise literature on gamified annotation systems and intensively engage with ingrained design artifacts, research activities, and design rationales to infer how they shape contemporary design knowledge on gamified annotation systems. We provide a broad overview of the literature and deep insights into 13 solution streams that describe existing design knowledge in the form of means-end relationships between goals and solutions (Maedche et al. 2019; Vom Brocke, Hevner, and Maedche 2020). We abstract our findings to derive recommendations for the creation of more mature design knowledge on gamified annotation systems.

The principal contribution of this study lies in a synthesis of existing gamified annotation system artifacts and the design knowledge therein. In doing so, we contribute to laying a base of useful knowledge (including descriptive and prescriptive knowledge) on designing gamified annotation systems. The ability to effectively leverage extant design knowledge is dependent on efficiency and cost of access to this knowledge (Gregor and Hevner 2013). To this end, we facilitate designers to effectively exploit extant design knowledge and thereby contribute to the creation of successful gamified annotation systems. We derive recommendations which can guide future research endeavours to purposively derive more abstract, complete, and mature design knowledge on gamified annotation systems. Our results are also interesting for gamification researchers. The derived solution streams showcase how gamification-based solutions can address certain design goals in light of the individual and situational factors of the AI training data annotation context. By investigating which factors support or inhibit the maturity of individual solution streams and tying these back to existing discourses in research on gamification design, we contribute to a better understanding of which discourses may be important in the quest to understand and support successful gamification design.

## 2. Foundations

### 2.1. AI training data annotation

Data annotation broadly describes the addition of metadata (often also called 'annotations' or 'labels') to existing data instances (Pustejovsky and Stubbs 2012). The most frequently annotated formats of data are images (Alvi 2024) and text documents (SuperAnnotate 2023). Less common data formats include audio files (Callaghan et al. 2018) and sensor data (L'Heureux et al. 2017). Regarding the added metadata, common forms include tags, labels, bounding boxes, or a combination of those (Pustejovsky and Stubbs 2012). Table 1 lists common example annotation tasks from extant

**Table 1.** Example AI training data annotation tasks for some common data formats.

| Data format | Added metadata + Example task |
|---|---|
| Images | Segmentation masks for anatomical structures in surgical images (Wagner et al. 2021) |
| | Tags that describe emotion in abstract images (Mekler et al. 2017) |
| Text | Markings for linguistic expressions that refer to same real-world entity (Kicikoglu et al. 2019) |
| | Adding Part-of speech tags to text documents (Pustejovsky and Stubbs 2012) |
| Audio | Label presence of abnormal heart sounds (Callaghan et al. 2018) |
| Sensor data | On-Off Labels for electronic devices (L'Heureux et al. 2017) |

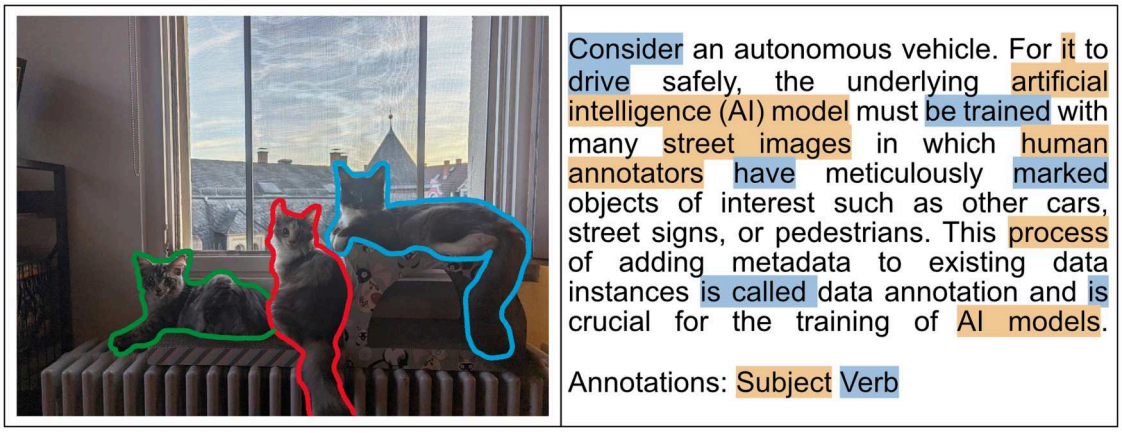


**Figure 1.** Example annotations.

literature grouped by data format, whereas Figure 1 shows two example annotations: an image of one of the authors' cats, annotated with segmentation masks, and the first three sentences of this paper's introduction, annotated with markings for subjects and verbs.

While data annotation can serve various purposes (e.g. organisation of personal data archives; Maltzahn et al. 2014), most annotation tasks today are performed to create AI training data (Meireles et al. 2021; Russo et al. 2021). Annotated training data allows AI models to easily recognise patterns and make inferences and is thus the primary way to train AI models (Khalaji et al. 2023). Effective data annotation is an integral process in the training and deployment of AI models (Gudivada, Apon, and Ding 2017). The primary goal of any annotation task is to create high-quality annotations (Warsinsky et al. 2022). What constitutes 'high-quality annotations' is context-dependent and subjective (Wand and Wang 1996). For the purpose of this study, we draw on the extensive discourse on data quality in the information systems (IS) domain and define annotation quality as 'fitness for use' (Wand and Wang 1996). This abstract view is sufficient to enable an informed discourse about increasing annotation quality through gamification without delving into the intricacies of data quality literature (for more information, we recommend Batini et al. (2009)).

Annotation research features several literature reviews which have compiled overviews of existing annotation tools in specific domains, such as computer vision (Sager, Janiesch, and Zschech 2021), biomedical imaging (Neves and Leser 2014) or document annotation (Neves and Ševa 2021). While these reviews give a good overview of existing annotation tools, gamification is rarely more than anecdotally mentioned (Neves and Leser 2014; Neves and Ševa 2021). Only the review by Sager, Janiesch, and Zschech (2021) highlights gamification as an image labelling support technique and explicitly investigates the presence of gamification in existing annotation tools for computer vision. Yet, they find no such occurrences, indicating that gamification is not supported by existing annotation tools specific to the computer vision domain. Compared to these reviews, our study covers a broad range of annotation domains and is not limited to specific domains.

### 2.2. Design knowledge

Broadly, design knowledge may be viewed as a means-end relationship between a problem space and a solution space (Maedche et al. 2019; Vom Brocke, Hevner, and Maedche 2020), thereby taking the form of 'to achieve <goal G>, use <solution S>'. Design knowledge may be broken down into two basic types (Gregor and Hevner 2013; Hevner et al. 2004): (1) descriptive knowledge, which is the 'what' knowledge that helps to analyze, explain, and predict phenomena, and (2) prescriptive knowledge, which is the "how' knowledge of human-built artifacts' (Gregor and Hevner 2013, 343). Building a mature design knowledge base in any domain requires both descriptive and prescriptive knowledge. While prescriptive knowledge provides insights into how artifacts affect their environment (e.g. individuals, organisations), descriptive knowledge 'enhance[s] our understanding of the world and the phenomena our technologies harness (or cause)' (Vom Brocke, Hevner, and Maedche 2020, 10–11).

In this work, to characterise extant design knowledge on gamified annotation systems, we embrace the form of design knowledge as a means-end relationship between problem and solution space (Maedche et al. 2019; Vom Brocke, Hevner, and Maedche 2020). To do so, we introduce the concept of solution streams. This is formalised in a conceptual framework (see Figure 2), associated with definitions of important concepts in Table 2. In general, our view is as follows: extant

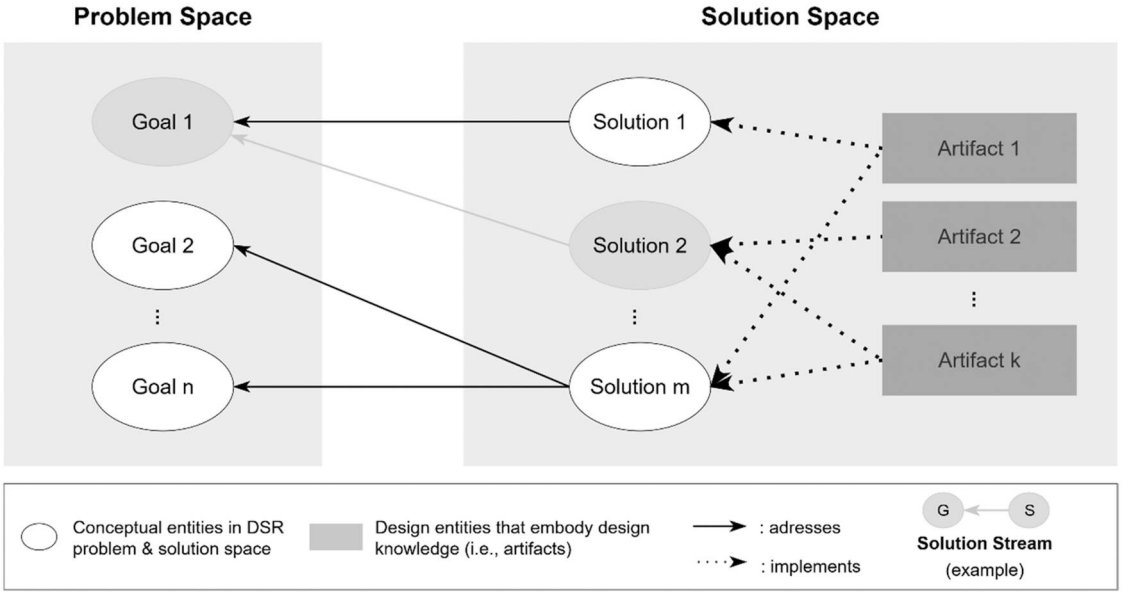


**Figure 2.** Conceptual framework of solution streams.

literature on gamified annotation systems features various artifacts, which implement one or more solutions, which in turn address one or more goals (i.e. ends). Goals are considered the key conceptual entity when conceptualising a problem space, as they mediate between abstract needs and concrete requirements (Maedche et al. 2019). Therefore, we use goals to represent the problem space. We then construe a solution stream to be a (goal, solution)-tuple, whereby the solution addresses the respective goal. Solution streams abstract away from individual artifacts as well as contextual or situational factors (Nacke and Deterding 2017), and rather represent pieces of design knowledge (as means-end relationships) for an overall system class (here, gamified annotation systems). Additionally, the maturity of each solution stream can be individually investigated to paint a granular picture of the current state of design knowledge. Therefore, we use solution streams to investigate the state of design knowledge on gamified annotation systems.

**Table 2.** Definitions of important concepts.

| Concept | Definition | Example |
|---|---|---|
| Goal | Desired results or a desired state of affairs (Maedche et al. 2019) | 'Improve user motivation' |
| Solution | Concrete, tangible features of an artifact (Strohmann and Khosrawi-Rad 2025) | 'provide rewards for task completion' |
| Artifact | Design entities (constructs, models, methods, or instantiations (Hevner et al. 2004)) that embody design knowledge (Strohmann and Khosrawi-Rad 2025). | An existing gamified annotation system instantiation |
| Solution Stream | A (goal, solution) tuple where the solution addresses the respective goal | ('Improve user motivation', 'provide rewards for task completion') |

To describe the maturity of design knowledge on gamified annotation systems in a comprehensible and informative way, we draw on two frameworks from the design science research (DSR) domain. First, we draw on the design research activities (DRA) framework (Maedche, Gregor, and Parsons 2021), which classifies studies into four quadrants based on (1) the type of knowledge contribution (i.e. whether descriptive or prescriptive knowledge is produced) and (2) the researcher role (i.e. whether a research team creates a new artifact or observes existing artifacts). By classifying studies from a stream of research into the respective quadrants, the DRA framework is useful to determine the maturity of extant design knowledge (Maedche, Gregor, and Parsons 2021). Doing so may, for example, unearth a lack of studies in a quadrant or a lack of mature design artifacts in a quadrant, which can provide insights into whether further work is warranted in specific quadrants.

Second, we draw on the framework of DSR contributions types (Gregor and Hevner 2013), which proposes that the types of artifacts that exist in a research domain can provide insights into the maturity of the respective design knowledge base. The framework distinguishes three levels of contribution: At the lowest level of contribution are instantiations (i.e. situated implementations of artifacts), while design theories sit at the highest level of contribution. The space in between is populated by mid-range artifacts such as models, methods or design principles, which provide more abstract and complete knowledge than instantiations,

yet not enough to be considered design theory. The basic idea is that offering artifacts on a more abstract level allows them to be operationalised and studied in several contexts, providing more complete and mature design knowledge (Gregor and Hevner 2013).

### 2.3. Designing gamified annotation systems

Gamification describes the use of game design elements like points, badges or leaderboards in non-game contexts (Deterding et al. 2011). In IS, gamification is often used as a purposive design strategy to evoke gameful experiences and motivate individuals to perform certain tasks (Koivisto and Hamari 2019). The principal design goal of a gamified IS is to support meaningful engagement: that is, to evoke both instrumental outcomes such as a behaviour change and experiential outcomes such as flow or cognitive absorption (Liu, Santhanam, and Webster 2017). To this end, gamification research has focused on creating descriptive knowledge that can explain or predict how gamification works and prescriptive knowledge that can provide actionable prescriptions for gamified IS design. Regarding descriptive knowledge, an important issue in gamification design is to choose those experiential outcomes that effectively lead to higher levels of desired instrumental outcomes, as these are highly context-dependent and thus vary based on factors like prospective users or characteristics of the task one seeks to gamify (Liu, Santhnam, and Webster 2017; Palmquist 2024). Regarding prescriptive knowledge, research has created several resources to inform the design of gamified IS, including gamification design frameworks that aim to provide actionable prescriptions for designing a gamified IS (e.g. by guiding the choice of gamification elements; Mora et al. 2017) and design principles that provide high-level guidance for the design of gamified IS overall (Liu, Santhnam, and Webster 2017; Morschheuser et al. 2018) or in specific contexts (e.g. adaptive gamification for online courses; Sezgin and Yüzer 2022).

We identify several literature reviews that synthesise extant design knowledge in domains related to gamified annotation. Most notably, we relate to the domain of gamified crowdsourcing (Yang, Ye, and Feng 2021), as annotation tasks are frequently crowdsourced. To this end, Morschheuser et al. (2017) systematically describe the landscape of gamified crowdsourcing, including aspects like gamification affordances, domains, or psychological outcomes, and subsequently derive recommendations for designing gamified crowdsourcing systems (e.g. to consider personal factors in the design process). However, annotation tasks are only one of several different crowdsourcing tasks, and the results do not extend to annotation tasks that are not crowdsourced (e.g. expert annotation; Warsinsky et al. 2022). Additionally, we relate to research on gamifying learning tasks, which sometimes feature annotation tasks to support learning (e.g. annotating words in a text to improve reading comprehension; C. M. Chen, Li, and Chen 2020). Studies in this domain have also recognised the lack of learner motivation as a core issue (Koivisto and Hamari 2019), and several systematic reviews have been conducted to synthesise extant knowledge in this area (Carmichael, MacEachen, and Archibald 2022; Ertan and Kocadere 2022). However, these existing reviews are mostly focused on understanding how gamification works in learning tasks (i.e. descriptive knowledge) and offer little design guidance. Additionally, when annotating for learning purposes, annotators (i.e. learners) usually directly benefit from performing the task via learning effects (C. M. Chen, Li, and Chen 2020), while AI training data annotation offers little direct benefits to annotators (Warsinsky et al. 2022).

In sum, while prior reviews provide valuable knowledge on gamification in adjacent domains like crowdsourcing, learning or annotation tools, we are not aware of any review that has investigated the design of gamified annotation systems as a distinct system class. To address this gap, our review synthesises design knowledge specific to gamified annotation systems; regardless of whether they are crowdsourced or embedded in learning contexts. Additionally, compared to existing reviews, we also explicitly evaluate the maturity of existing design knowledge in terms of individual solution streams, which we synthesise into explicit recommendations for the creation of more mature design knowledge for gamified annotation systems.

## 3. Literature review approach

### 3.1. Data collection

To answer our research question, we conducted a semi-systematic review (Snyder 2019). We deemed this type of review suitable for our work, as it is designed for topics that have been studied by diverse groups of scholars in various disciplines, which fits with gamified annotation being employed in various disciplines (e.g. healthcare; Dumitrache et al. (2013); or cultural heritage preservation; Toumanidis et al. (2019)). We combine a systematic database search with qualitative analysis methods to synthesise the state of design knowledge on gamified annotation systems (Snyder 2019). Our data collection approach is summarised in Figure 3. We employed a systematic search strategy following the guidelines by Levy and Ellis (2006). We opted for

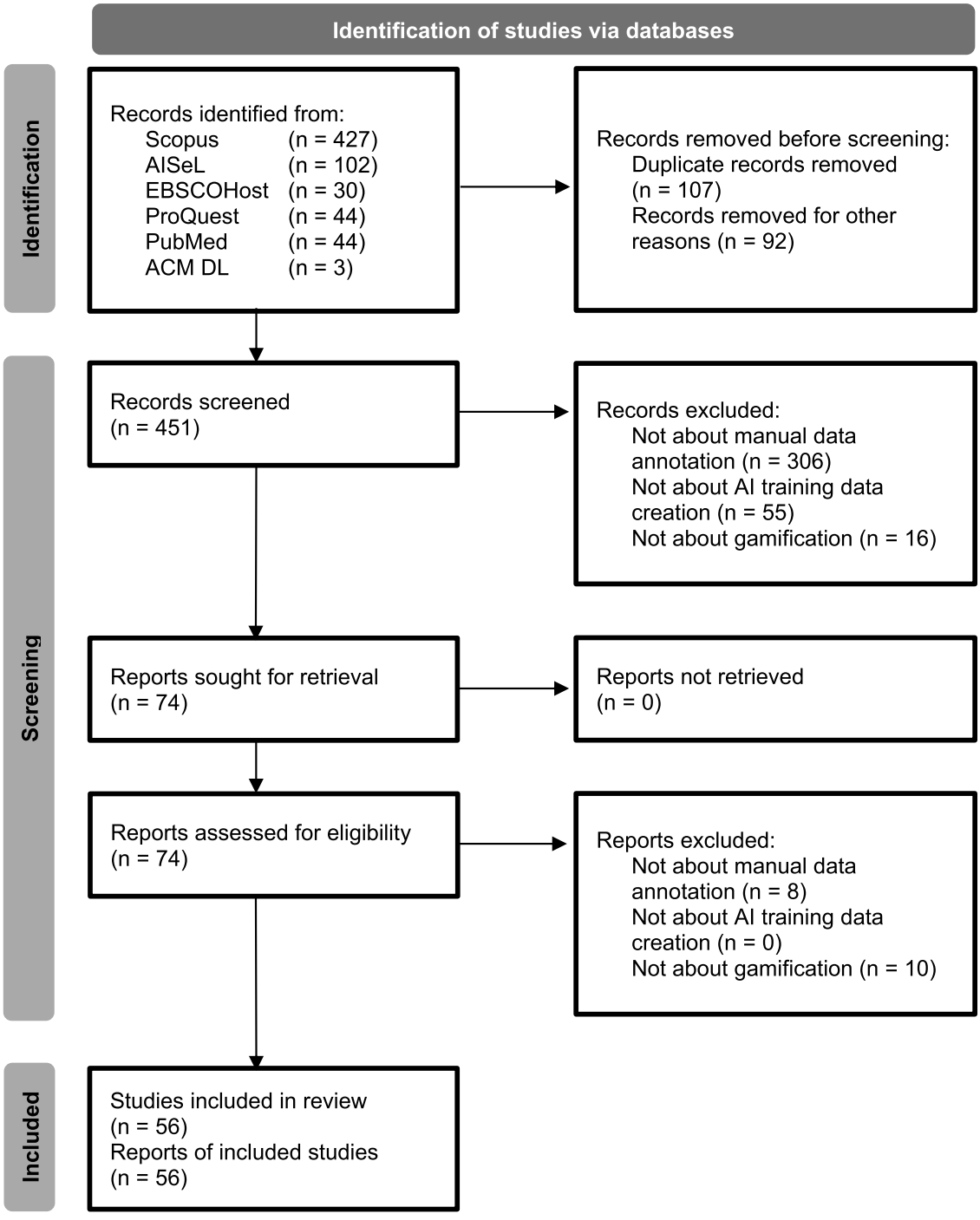


**Figure 3.** PRISMA flow diagram visualising our data collection approach (adapted from Page et al. 2021).

a combination of databases with broad, multidisciplinary coverage (i.e. Scopus, EBSCOHost, and ProQuest), focused databases that contain literature on gamification and human–computer interaction (i.e. ACM Digital Library and AISeL), and a database on healthcare (i.e. PubMed), as healthcare is a common application domain of annotation. We searched in title, keywords, and abstract with a search string that covered key terms related to gamification and annotation (see Table 3). The search string also included terms that are often used synonymously or interchangeably with annotation, such as labelling (Rädsch et al. 2023),

**Table 3.** Search string for our systematic database search.

| Key terms | Search string including alternative terms |
|---|---|
| Gamification | gamif* |
| | AND |
| Annotation | (annotat* OR label* OR segment* OR tag*) |

segmentation (Wang et al. 2021), or tagging (Pustejovsky and Stubbs 2012). Our search on January 9th, 2024, yielded 650 publications.

We excluded 107 duplicates, 23 articles that were not in English or not peer-reviewed, and 69 conference proceedings (i.e. we excluded the conference proceedings as works themselves but included individual conference papers). We then used predefined exclusion criteria to assess the relevance of the remaining 451 articles (see Table 4). To increase the validity of the assessment, two authors independently screened the abstract of each article and discussed conflicting assessments until all discrepancies were resolved. To account for inter-rater reliability, we also report Cohen's Kappa (κ; Cohen (1960)).

From the 451 studies assessed, we excluded 314 articles that did not involve manual data annotation (intercoder agreement = 87.14% and $\kappa = 0.7428$). Following the definition of annotation as the 'addition of metadata to existing data instances' (Pustejovsky and Stubbs 2012), we screened each abstract in search for the existing data instances and the respective metadata added. We posited that the metadata should have a clear relation to the associated data instance yet should not provide meaningful value on its own (i.e. without the original data instances being annotated). We did so in particular to exclude transcription or translation tasks, which are usually not done with AI training in mind. Furthermore, we excluded papers that deal with purely automated annotation systems, such as systems with ML models that annotate without any human intervention (Washington et al. 2022). In the next step, we excluded 55 papers (intercoder agreement = 77.37% and $\kappa = 0.5474$) that explicitly stated an annotation purpose other than creating AI training data (e.g. annotation of a personal photo library for organisation purposes; Maltzahn et al. (2014)). Lastly, we excluded 26 articles that did not describe a gamified system (intercoder agreement = 80.49% and $\kappa = 0.6098$). To differentiate between gamification and often synonymously used terms (most notably, serious games and games-with-a-purpose), we drew on the definition of gamification as use of game design elements in non-game contexts (Deterding et al. 2011) and accordingly assessed whether the described systems can be considered full games and included only those which encompassed a removable game layer (Warsinsky et al. 2021). Ultimately, our data collection yielded a set of 56 relevant articles.

**Table 4.** Summary of the exclusion criteria applied during screening and full-text assessment.

| Exclusion criteria | Description |
|---|---|
| Not about manual data annotation | The publication does not describe an artifact intended for manual addition of metadata to existing data instances |
| Not about AI training data creation | The publication explicitly mentions an annotation purpose other than AI training data creation |
| Not about gamification | The publication does not describe an artifact with gamification elements encompassed in a removable game layer |

### 3.2. Data analysis

To analyze our 56 relevant articles, we followed a manual concept-centric data analysis approach informed by Webster and Watson (2002). We engaged with the full text of each paper and coded them based on predefined units of analysis (see Table 5), which comprised general information about the studies (e.g. applied method and domain), annotation-related data (e.g. annotated data and added metadata), and gamification-related data (e.g. instrumental outcomes and experiential outcomes). We compiled a concept matrix (Webster and Watson 2002) of the coded papers and formed the coded information into frequency tables, which form the first part of our results (see section 4.1).

The most important unit of analysis was the included design knowledge in the papers. We coded for the types of artifacts presented in the papers (Gregor and Hevner 2013) and investigated in particular those artifacts which were more abstract than instantiations. We also investigated the design research activities in each study and assigned it into its respective quadrant in the DRA framework (Maedche, Gregor, and Parsons 2021). After doing so, we followed the method suggested by Maedche, Gregor, and Parsons (2021) and investigated the individual quadrants to gain an overview of the research activities in that literature stream. These insights form the second part of our results (see sections 4.2 & 4.3).

To gain a deeper and more granular look into the presented gamified annotation system designs, we carefully analyzed each paper and coded for text passages that featured design rationales where the authors explain their design decisions (e.g. why specific gamification elements are used). After coding these design rationales for each paper, we aimed to heighten our level of abstraction by identifying similarities, grouping similar design rationales and finding overarching topics. In doing so, we applied

**Table 5.** Coding dimensions.

| Category | Units of analysis |
|---|---|
| General information | Method, sample sizes, domain |
| Annotation-related | Annotated data, added metadata, annotators |
| Gamification-related | Gamification elements & mechanics, targeted instrumental & experiential outcomes and effects of gamification on them |
| Design knowledge-related | Type of presented artifact (Gregor and Hevner 2013), DRA framework quadrants (Maedche, Gregor, and Parsons 2021), design rationales |

**Table 6.** Solution stream maturity criteria.

| Solution stream maturity criterion | Example contributors (+) & inhibitors (-) |
|---|---|
| Presence of abstract artifacts in domain | + Abstract design artifacts exist to inform design |
| | + Existing abstract artifacts are specific to solution |
| | – Abstract design artifacts lack; most artifacts are instantiations |
| | – Existing abstract artifacts are only tangentially related to solution |
| Design research activities | + Artifacts are re-used across multiple studies |
| | + Studies create a balance of descriptive and prescriptive knowledge |
| | – Artifacts are not or rarely re-used; most studies construct new artifacts |
| | – Studies overwhelmingly create descriptive (prescriptive) knowledge |
| Presence of justificatory knowledge | + Studies use behavioural theories to explain why designs work the way they do |
| | + Studies provide detailed design rationales on individual components of their artifacts |
| | – Studies feature no design rationales |
| | – Studies feature mostly abstract design rationales |
| Knowledge on effectiveness of solution | + Empirical results specific to solution exist |
| | + Empirical results on solution are mostly positive |
| | – Little or no empirical results specific to solution |
| | – Empirical results on solution are mixed; with potential unintended negative side effects |
| Implementability | + Solution is intuitive to implement through gamification elements |
| | + Studies provide actionable guidance on how to implement solution through gamification elements |
| | – Solution is not intuitive to implement through gamification elements |
| | – Studies provide only abstract guidance how to implement solution through gamification elements |

our conceptual model and synthesised solution streams that connect relevant design goals (i.e. ends) with pertinent gamification-based solutions (i.e. means). Ultimately, we identified 13 such solution streams.

For each solution stream, we gauged the maturity of existing design knowledge based on several criteria, which we summarise in Table 6. From a DSR perspective, important criteria were the presence of abstract artifacts (Gregor and Hevner 2013) and design research activities (Maedche, Gregor, and Parsons 2021). From a gamification perspective, important criteria were knowledge on the effectiveness of a solution (i.e. 'does it work?'; Hamari, Koivisto, and Sarsa (2014)), the presence of justificatory knowledge that helps to explain why a solution works the way it does (Gregor and Jones 2007; Koivisto and Hamari 2019; Nacke and Deterding 2017), and the implementability of the solution (Iivari, Rotvit Perlt Hansen, and Haj-Bolouri 2021). For each solution stream, we assessed these criteria and derived an overall evaluation of its maturity as either low, medium or high. These insights form the second major part of our results (see section 5).

## 4. Overview of analyzed literature

### 4.1. Methodical approaches and annotation tasks

We begin with a high-level overview of the 56 investigated papers. Table 7 provides an overview of the reviewed studies' methodological approaches (Table 7(a)), annotation domains (Table 7(b)), and annotators (Table 7(c)). Most studies applied quantitative approaches (n = 34), with the most frequently applied individual method being experiments (n = 18). The annotation domains of the reviewed literature were quite spread, with prominent domains being crowdsourcing (n = 10), linguistics (n = 9), and medicine (n = 7). Regarding annotators, most studies (n = 34) focused on a crowd as annotators, while 15 studies did not feature a specific group of annotators. Two studies focused specifically on expert annotators (Sevastjanova et al. 2021; Warsinsky et al. 2022), while two other studies contrasted expert annotators with crowd annotators (Ivanjko 2019; Öhman et al. 2018). Two studies (Anagnostou, Mollas, and Tsoumakas 2018; Kicikoglu et al. 2019) investigated users of

**Table 7.** Types of studies, domains, and annotators in our reviewed literature.

| Type of study | # of studies | Domain | # of studies | Annotators | # of studies |
|---|---|---|---|---|---|
| Conceptual | 9 | Crowdsourcing | 10 | Crowd | 34 |
| Quantitative | 34 | Linguistics | 9 | None specific | 15 |
| Qualitative | 13 | None specific | 9 | Company employees | 2 |
| | | Medicine | 7 | Embedded system users | 2 |
| | | Speech processing | 5 | Experts | 2 |
| | | Activity recognition | 5 | Crowd vs experts | 2 |
| | | Cultural heritage | 3 | | |
| | | Work, Social networking | 2 | | |
| | | Software development, Translation, Sensor data, Information retrieval | 1 | | |
| (a) | | (b) | | (c) | |

platforms in which they embedded annotation tasks (e.g. the YouTube platform; Anagnostou, Mollas, and Tsoumakas (2018)). Similarly, two studies (Alaghbari et al. 2020, 2021) investigated employees of a company where they gamified annotation processes.

Annotation tasks in the reviewed literature were varied; Table 8 provides an overview of annotation tasks we encountered, grouped by the format of annotated data. Note that one study featured annotation of multiple data formats (Toumanidis et al. 2019) and several studies featured different kinds of added metadata. We mostly encountered annotations of image or video data (n = 26), such as abstract art images (Lessel et al. 2019; Lessel et al. 2022) or images of skin lesions (Balducci and Buono 2018; Duhaime et al. 2023), closely followed by textual data (n = 18), such as scanned documents (Alaghbari et al. 2020, 2021). Other data formats like audio (Hantke, Appel, and Schuller 2018) or sensor data (Mairittha and Inoue 2018; Mairittha, Mairittha, and Inoue 2019) were less present in the reviewed literature. Across all data formats, the most popular added metadata were tags, which could be either binary (e.g. translation problem flags; Y. M. Chen 2019), chosen from a predefined set of tags (e.g. a set of unique sentiments and emotions; Öhman and Kajava (2018); Öhman et al. (2018)), or freely added by the annotators (e.g. moods conveyed in images; Mekler et al. (2013a); Mekler et al. (2013b); Mekler et al. (2017)).

### 4.2. Existing artifacts

Investigating the types of artifacts presented in the reviewed literature (for an overview, see Table 9) revealed that most studies featured instantiations (n = 45), while we identified eleven artifacts that are more abstract. Two studies develop a model and use the prescriptive knowledge therein to subsequently build an instantiation (Dumitrache et al. 2013; Mairittha et al. 2021). Two abstract artifacts had little relation to the gamification-part of the presented gamified annotation systems. Rather, these artifacts include a method to make annotated datasets available to the public (Cao et al. 2015), and a method to create a platform that 'provides enhanced services and enables human–computer collaboration for data annotations and enrichment' (Chortaras et al. 2018, 1117). The remaining artifacts are somehow related to individual gamification-based solutions. Hence, we describe them in more detail in the solution streams section when their respective content is relevant.

**Table 9.** Artifact types in our reviewed literature.

| Artifact type | # of studies |
|---|---|
| Instantiation | 45 studies |
| Method | 4 studies (Cao et al. 2015; Chortaras et al. 2018; Gurav, Parkar, and Kharwar 2020; Kicikoglu et al. 2019) |
| Model | 5 studies (Chamberlain, 2014; Chen 2019; Friðriksdóttir and Einarsson 2022; Jin et al. 2020; L'Heureux et al. 2017) |
| Model + Instantiation | 2 studies (Dumitrache et al. 2013; Mairittha et al. 2021) |

### 4.3. Design research activities

Figure 4 features an overview of how the reviewed literature is situated in the DRA framework quadrants (Maedche, Gregor, and Parsons 2021). Investigating each quadrant more in-depth revealed the following insights:

**Construction Quadrant.** This quadrant hosted most of the reviewed studies. Studies in this quadrant involved the creation of a new gamified annotation system instantiation, where the gamification part of the

**Table 8.** Overview of annotation tasks.

| Annotated data | # of studies | Added metadata | # of studies | Example annotations |
|---|---|---|---|---|
| Images & Video | 26 | Tags | 17 | Tags to convey mood in abstract art images (Mekler et al. 2017) |
| | | Segmentation masks | 6 | Segmentation of vessels in 3D vascular images of airways (Huang and Hamarneh 2017) |
| | | Bounding boxes | 2 | Bounding boxes around cultural heritage sites (Toumanidis et al. 2019) |
| | | Labels | 3 | Natural language descriptors for body movements (Plappert, Mandery, and Asfour 2016) |
| | | Non-specific | 1 | (Jin et al. 2020) |
| Text | 17 | Tags | 11 | 'Translation problem flags' (Chen 2019) |
| | | Labels | 6 | Labels around multi-word expressions (Fort et al. 2020) |
| | | Bounding boxes | 2 | Adding bounding boxes around specific parts of text documents (Alaghbari et al. 2020, 2021) |
| Audio | 7 | Tags | 7 | Tags for emotion conveyed in speech (Hantke et al. 2015) |
| Sensor data | 5 | Labels | 3 | Adding activity labels to smartphone accelerometer data (Mairittha et al. 2021) |
| | | Tags | 2 | Adding on-off tags for electronic devices (Cao et al. 2015; L'Heureux et al. 2017) |
| Non-specific | 3 | - | - | - |

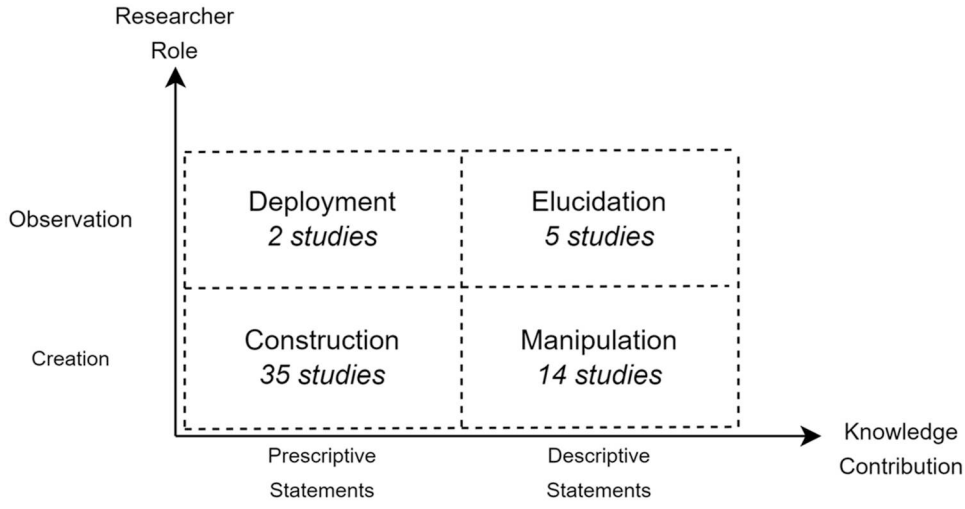


**Figure 4.** Overview of design research activities following the DRA framework (Maedche, Gregor, and Parsons 2021).

system was sometimes quite substantial and sometimes only a minor component of the resulting instantiation (Anagnostou, Mollas, and Tsoumakas 2018; Plappert, Mandery, and Asfour 2016). Likewise, provided rationales for gamification design decisions varied greatly in depth, with some studies meticulously justifying each gamification element individually (Hantke et al. 2015; Hantke, Appel, and Schuller 2018; Sevastjanova et al. 2021), and other studies providing quite high-level rationales such as increasing engagement with the annotation task (Anagnostou, Mollas, and Tsoumakas 2018). Some studies provide only the description of the created artifact (Alaghbari et al. 2020, 2021; Balducci and Buono 2018), while others include evaluation activities which involved mostly low sample sizes and qualitative approaches to gain initial insights on the created artifacts (Guillot et al. 2016; Hantke et al. 2018).

**Deployment Quadrant.** We identified two studies in the deployment quadrant. Both these studies use existing gamified annotation system architectures to launch their annotation tasks, and use the resulting instantiations to collect evaluation data in the context of a specific text annotation task (i.e. noun-noun compound annotation; Bos and Nissim (2015)) or to compare performance metrics between non-expert and expert annotators (Duhaime et al. 2023).

**Manipulation Quadrant.** The manipulation quadrant consists almost exclusively of experiment studies, where existing gamified annotation system instantiations are adapted to create descriptive knowledge about specific gamification-related aspects. Manipulations include the presence or nature of certain gamification elements (Mekler et al. 2013a; Mekler et al. 2013b; Mekler et al. 2017), the level of customisation of gamification elements (Lessel et al. 2019; Schubhan et al. 2020), or personalisation based on user types (Altmeyer et al. 2022). Importantly, studies that fall under this category often acknowledge that they do not necessarily seek to build the most effective annotation system (Mekler et al. 2017), but rather use annotation as a research context for gamification to ensure replicability and comparability of results, or to have easy-to-measure outcomes (Altmeyer et al. 2022; Lessel et al. 2022; Mekler et al. 2017). Several studies in this quadrant were based on the same initial instantiations. Mekler et al. (2013b) created a gamified annotation platform, which was subsequently creatively adapted for experiments by Mekler and colleagues themselves (Mekler et al. 2013a; Mekler et al. 2017) and by other research teams (Altmeyer et al. 2022; Lessel et al. 2022; Schubhan et al. 2020). Similarly, Mairittha and Inoue (2018) developed an instantiation and later manipulated it to conduct experiments on it in further studies (Mairittha et al. 2021; Mairittha, Mairittha, and Inoue 2019).

**Elucidation Quadrant.** The elucidation quadrant includes mostly studies that continue the efforts from the manipulation quadrant by continuing to observe the created artifacts (Hantke et al. 2019; Hantke, Zhang, and Schuller 2017) and producing descriptive knowledge on specific aspects of the artifacts, like the customisation of gamification elements (Lessel et al. 2019) or annotators' positioning on a leaderboard (Na and Han 2023). Ivanjko (2019) uses an existing gamified annotation system to compare expert and non-expert annotators. Lastly, one study did a case study to provide a detailed breakdown of an existing gamified annotation system (Y. M. Chen 2019) and one study that did a literature review to provide a broad overview of artifacts for crowdsourcing annotation (Jin et al. 2020).

**Summary.** Our analysis with the DRA framework indicates that in most of the reviewed studies, researchers create new artifacts and investigate them, as signified by the high number of studies in the construction and manipulation quadrants. Noticeably, only the gamified abstract art image annotation platform by Mekler et al. (2013b), was applied in studies by researchers outside of the initial research team that created the artifact (Lessel et al. 2022; Na and Han 2023; Schubhan et al. 2020). Regarding studies beyond the construction quadrant, there is a low number of studies in quadrants where existing artifacts are observed (i.e. deployment, elucidation), but several studies in the manipulation quadrant that provide insightful descriptive knowledge.

## 5. Solution streams in current research on gamified annotation systems

By engaging with existing design rationales for gamified annotation systems artifacts, we synthesised 13 solution streams. Table 10 features an overview of the identified solution streams, along with our overall assessment of whether each individual stream's maturity is low, medium or high, and those criteria we deemed the most decisive for each stream's maturity (an assessment of all criteria can be found in Appendix A). In the following, we describe each solution stream, ordered by goals for readability.

### 5.1. Goal 1: ensure that annotators feel compensated for their work

Annotation tasks on their own commonly offer little inherent rewards or incentives for annotators to perform them. Annotators often get no feedback on the quality of their work and are not the ones that will be using the produced annotated data themselves. Therefore, an important goal is to ensure that annotators feel compensated for their work.

*Solution Stream 1: To ensure that annotators feel compensated for their work, provide incentives that support intrinsic motivation.* Drawing on self-determination theory, intrinsic motivation describes doing a task driven by inherent satisfactions of the task itself and not by external rewards (Ryan and Deci 2020). In gamified annotation systems, achieving intrinsic motivation is considered an important success criterion (L'Heureux et al. 2017) that is especially important when financial (i.e. extrinsic) incentives lack (Hantke et al. 2015; Jin et al. 2020). Therefore, studies frequently include incentives that support intrinsic motivation. One such incentive is self-learning, where gamification features provide annotators with ways to learn and improve their skills (Duhaime et al. 2023), for example through a feedback dashboard that allows annotators to reflect on their previous annotations (Gutiérrez Páez et al. 2021). The reviewed literature proposes that incentives such as self-learning can support intrinsic motivation, which allows annotators to feel compensated for their work. However, evoking intrinsic motivation is difficult: several studies present gamified annotation system instantiations that failed to evoke intrinsic motivation (Mekler et al. 2013a; Mekler et al. 2013b). Moreover, it is inconclusive whether aiming to achieve intrinsic motivation is really expedient: Several studies report positive effects of gamification on instrumental outcomes (e.g. annotation quantity; Mekler et al. (2013a); Mekler et al. (2013b)), despite no effects on intrinsic motivation. Other studies report positive effects such as increased engagement when using extrinsic incentives (Feyisetan et al. 2015). Ultimately, designing for intrinsic motivation in gamified annotation systems appears difficult, and we lack knowledge on how to design incentives that are effective in evoking intrinsic motivation and can transcend extrinsic incentives. Thus, we consider this a low maturity solution stream.

*Solution Stream 2: To ensure that annotators feel compensated for their work, provide retrospective rewards if ground truth is not clear at time of annotation.* When annotators finish annotating a data instance or a dataset, they often expect some form of compensation, otherwise they might get frustrated (Chamberlain, 2014; Kicikoglu et al. 2019). Yet, providing adequate compensation at the time of annotating can be difficult, as the quality of an annotation is often not clear, for example if there are no other annotations to compare it to (Kicikoglu et al. 2019). To this end, gamification elements can be designed as retrospective rewards: when finishing an annotation, annotators may receive a token amount of points as immediate compensation, plus a digital egg, which then hatches and provides points once the quality of the annotation can be assessed (Kicikoglu et al. 2019). In genreal, studies that applied this solution reported positive effects from their instantiations (Kicikoglu et al. 2019; Sevastjanova et al. 2021). Regarding design guidance, Chamberlain (2014) presents an 'Annotation-Validation Model' that can guide the design of rewards through retrospective validation specifically when ground truth annotations are not available. Overall, retrospective rewards through gamification elements appear to be an effective solution to tackle the frequent issue of lacking ground truth, with ample design guidance. We thus consider this solution stream to be of high maturity.

**Table 10.** Overview of gamified annotation system solution streams.

| Goal | # | Gamification-based solution | Solution maturity | Contributors (+) & inhibitors (-) to solution maturity |
|---|---|---|---|---|
| Ensure that annotators feel compensated for their work | 1 | Provide incentives that support intrinsic motivation | Low | + Self-determination theory as an asset to inform design<br>– Intrinsic motivation is difficult to design for<br>– Inconclusive whether aiming to achieve intrinsic motivation in gamified annotation systems is really expedient |
| | 2 | Provide retrospective rewards if ground truth is not clear at time of annotation | High | + Effective to circumvent issues from lack of ground truth annotations<br>+ Existing model to inform design of rewards through retrospective validation specifically (Chamberlain, 2014) |
| Ensure that annotators perceive value from using the annotation system | 3 | Provide meaningful framing of annotation task | High | + Intuitive to implement by embedding annotation task into narrative<br>+ Strong empirical results for effectiveness from studies in manipulation quadrant |
| | 4 | Hide annotation task behind game | Med | + Can effectively shift system focus to entertainment, providing primarily hedonic value to annotators<br>– Unclear where to settle instantiations on game continuum |
| | 5 | Create meaningful social interactions | Med | + High potential to support meaningful engagement<br>– Several possible negative side effects; lack of design guidance on how to avoid or tackle these |
| Support continuous engagement | 6 | Provide variety in how to interact with the annotation system | High | + Almost all gamification elements implicitly provide alternative ways to interact with an annotation system<br>+ Existing models to support design for variety specifically (Kicikoglu et al. 2019) |
| | 7 | Gradually raise task difficulty | Low | + Flow theory as an asset to inform design<br>– Unclear how to apply solution if ground truth annotations are not present to assess difficulty of annotating specific data instances |
| | 8 | Show annotators their annotation progress | High | + Subtle solution that can be realised through various gamification elements<br>+ No abstract artifacts; but ample design inspiration from existing instantiations with sophisticated progress systems |
| | 9 | Adapt to individual differences in annotators | Low | + Several existing user type models to guide design<br>– Effectiveness of solution unclear; many neutral or mixed empirical results<br>– Importance of solution questionable; unclear when worth the implementation effort |
| Shape desirable annotation behaviours directly | 10 | Reinforce (punish) annotation behaviours that contribute to increased (decreased) annotation quality | Med | + Effective to provide feedback to annotators; punishment can also tackle cheating behaviour<br>+ Intuitive to implement through points systems<br>– Punishment can lead to several negative effects; unclear how to avoid these negative effects |
| | 11 | Encourage timely annotation | Med | + Intuitive to couple gamification elements with annotation quantity<br>+ Fairly large empirical support for effectiveness<br>– Unclear how to avoid jeopardising annotation quality while encouraging annotation quantity |
| Save resources | 12 | Replace monetary rewards with gamified rewards | Med | + Easy to implement<br>+ Existing model to inform design (Jin et al. 2020)<br>– Unclear how to design gamified rewards that are on par with monetary rewards |
| | 13 | Enable crowdsourcing instead of relying on expert annotators | Med | + Highly prominent to use gamification when crowdsourcing annotation tasks<br>+ Several existing models to guide creation of gamified crowdsourcing applications (Duhaime et al. 2023; Friðriksdóttir and Einarsson 2022; L'Heureux et al. 2017)<br>– Gamification design rarely focus of studies<br>– Specific mechanisms how gamification enables crowdsourcing remain unclear |

### *5.2. Goal 2: ensure that annotators perceive value from using the annotation system*

The value of an annotation task (e.g. improving AI models in a certain domain) is often not immediately visible to annotators (Mekler et al. 2013a). If annotators do not perceive value from doing the annotation task, they may be less interested in providing high-quality annotations, and more in exerting as little effort as possible. Therefore, an important goal is to ensure that annotators perceive some form of value from interacting with an annotation system.

*Solution Stream 3. To ensure that annotators perceive value from using the annotation system, provide meaningful framing of the annotation task.* Gamification elements can provide meaningful framing for an annotation task, most prominently by embedding the task

into a narrative (Mekler et al. 2013a; Toumanidis et al. 2019) which highlights the collective project goal to the annotators (e.g. improving AI in a certain domain; Gutiérrez Páez et al. (2021)). Doing so can help annotators gain a sense of purpose (Warsinsky et al. 2022) and provide value in particular to altruistic annotators as they gain the sense that they are contributing to a greater goal (Alaghbari et al. 2020, 2021). Empirical results support that meaningful framing can increase meaningful engagement in gamified annotation systems (Gutiérrez Páez et al. 2021). In particular, we found one expressive study in the manipulation quadrant by Mekler et al. (2013a), which found that when annotators were informed through a narrative how their annotation would improve science, they provided more sensible tags compared to those that were not. Overall, meaningful framing is effectively implementable through narratives and has strong empirical support that it is effective in allowing annotators to perceive value, making this a highly mature solution stream.

*Solution Stream 4: To ensure that annotators perceive value from using the annotation system, hide the annotation task behind game elements.* As annotation tasks are usually considered tedious, annotation systems are associated with work or labour (Friðriksdóttir and Einarsson 2022). Gamification elements can shift the perception of an annotation system to be primarily entertainment-focused (Friðriksdóttir and Einarsson 2022; Gurav, Parkar, and Kharwar 2020), which can divert annotators' awareness that they are doing an annotation task (Millour and Fort 2018) and thus make the task more compelling and fun (Chortaras et al. 2018). This entertainment from game elements can then provide hedonic value to annotators (Friðriksdóttir and Einarsson 2022; Kicikoglu et al. 2019). Some instantiations even go as far as trying to obfuscate from annotators that they are performing an annotation task altogether. For example, annotators may perceive taking images of geolocations not as annotation, but rather 'as a way to capture magical creatures, collect [virtual currency] and compete against others' (Simões and De Amicis 2016, 3). With regard to this solution stream, it is important to recognise that instantiations exist on a game continuum: while some only include lightweight gamification approaches (Millour and Fort 2018), others err more towards a full game (Simões and De Amicis 2016). Positioning on this continuum is an important design decision: for example, a more full game approach allows for easier hiding of the annotation purpose (Simões and De Amicis 2016), but may heighten annotators' expectations regarding the visual fidelity of game elements, as full games become the relevant benchmark (Hantke, Appel, and Schuller 2018; Mekler et al. 2017). Overall, while this solution seems effective, we lack design guidance that can inform the positioning on the described game continuum. Therefore, we consider this solution stream to be of medium maturity.

*Solution Stream 5: To ensure that annotators perceive value from using the annotation system, create meaningful social interactions.* As annotation tasks are mostly done in teams, a way to create perceived value for annotators is to create meaningful social interactions (Dumitrache et al. 2013), for example through team statistics (Guillot et al. 2016) or leaderboards (Öhman et al. 2018). Annotators seeing each other's work can give a sense of visibility (Alaghbari et al. 2020, 2021) and serve as a status symbol (Y. M. Chen 2019; Viana and Pinto 2017) or allow everyone to feel useful in reaching a bigger goal (Balducci and Buono 2018). Empirical findings suggest that when annotators look at the scores of their peers, they annotate more (Dumitrache et al. 2013), and may be motivated to perform as well as their peers (Jin et al. 2020; Viana and Pinto 2017). However, social comparison may also cause negative effects: newly joining annotators may get frustrated when comparing to annotators that have collected scores over a long time (Jauer, Spicher, and Deserno 2021; Jin et al. 2020; Warsinsky et al. 2022), and annotators that have reached the top of a leaderboard may become complacent (Na and Han 2023). Moreover, if annotation quality is measured based on agreement metrics (as is commonly the case; Bos and Nissim (2015); Öhman and Kajava (2018); Viana and Pinto (2017)), annotators could over rely on answers given by their peers (Dumitrache et al. 2013) and thus be inclined to seek the 'most agreeing' judgment as opposed to the correct one (Megorskaya, Kukushkin, and Serdyukov 2015). Ultimately, while social interactions are a promising solution, we still lack effective design guidance on how to avoid the several potential negative effects associated with it. Therefore, we consider this solution stream to be of medium maturity.

### 5.3. Goal 3: support continuous engagement

The repetitive and hardly-challenging nature of annotation tasks can make it difficult for annotators to stay engaged with the task, causing them to get easily distracted, which can lead to careless or sloppy mistakes. Annotators may also disengage with the annotation task altogether, leading to shorter annotation sessions (Na and Han 2023). Therefore, an important goal in gamified annotation systems is to support continuous engagement with the annotation task.

***Solution Stream 6: To support continuous engagement, provide variety in how to interact with the***

***annotation system.*** By design, annotators go through the same annotation process for lots of data instances, which can be repetitive and thus lead to mental boredom (Hantke, Zhang, and Schuller 2017; Mairittha et al. 2021). Gamification elements can provide alternative ways to interact with an annotation system (Cao et al. 2015; Hantke, Appel, and Schuller 2018; Öhman and Kajava 2018), for example by allowing annotators to play minigames (Warsinsky et al. 2022). Providing such variety can increase engagement (Mairittha et al. 2021) and contribute to longer annotation sessions (Na and Han 2023). One caveat to this solution is that to not disrupt annotation workflows (Mairittha, Mairittha, and Inoue 2019), it is important to identify opportune moments for annotators to engage with gamification elements (e.g. between annotation of individual data instances; Warsinsky et al. (2022)). Regarding design guidance, almost all gamification elements inevitably offer some degree of variety in system interactions, hence almost all gamified annotation system instantiations implicitly include this solution. If aiming for variety specifically, Kicikoglu et al. (2019) present the 'Motivation-Annotation Paradigm', which proposes to alternate between a motivation phase where the sole aim is for annotators to have fun, and an annotation phase, where the actual annotations are collected. Overall, providing variety has shown to be an effective solution to support continuous engagement, with ample design guidance, which is why we consider this a highly mature solution stream.

***Solution Stream 7: To support continuous engagement, gradually raise task difficulty***. In annotation tasks, the skill of the annotator typically exceeds the challenge of the task (Dumitrache et al. 2013; Warsinsky et al. 2022), which can lead to boredom and less engagement (Mairittha et al. 2021). Literature draws on flow theory (Nakamura and Csikszentmihalyi 2009) to argue that one solution to this is to gradually increase the task difficulty (Dumitrache et al. 2013; Riegler et al. 2015). This was most prominently implemented through points systems, where annotating more difficult data instances yields more points (Dumitrache et al. 2013; L'Heureux et al. 2017), but also by introducing time pressure (Vecchio et al. 2020) or progressively more difficult levels (Plappert, Mandery, and Asfour 2016; Riegler et al. 2015). For some studies it seemed easy to measure the difficulty of annotating individual data instances (e.g. Hantke et al. 2015; Riegler et al. 2015). However, those studies that did not feature ground truth annotations remarked on the struggle of indicating difficulty (Viana and Pinto 2017), and had to come up with sophisticated formalisations of difficulty scores (Dumitrache et al. 2013) or resort to manual allocation of points to data instances (L'Heureux et al. 2017). Therefore, gradually increasing task difficulty only appears to be a useful solution if ground truth data is present, which is however not the case in most annotation tasks. Thus, while conceptually effective in tackling boredom, we lack insights how to increase task difficulty in annotation tasks where ground truth data is not present. Therefore, we consider this a low maturity solution stream.

***Solution Stream 8: To support continuous engagement, show annotators their annotation progress.*** Annotators often have an inherent desire to know how much they have done (Plappert, Mandery, and Asfour 2016). Thus, fostering a sense of progression may motivate continued efforts (Mairittha et al. 2021) and nudge annotators toward completion of a task (Feyisetan et al. 2015). Many gamification elements lend themselves well to portray progress: this solution stream includes many studies situated in the construction quadrant, which feature various gamification elements to implement progress systems, including progress bars (Alaghbari et al. 2020; Feyisetan et al. 2015), points (Hantke, Appel, and Schuller 2018; Mairittha et al. 2021), trophies (Guillot et al. 2016) or levelling systems (Plappert, Mandery, and Asfour 2016). Evaluations of the created gamified annotation system instantiations are largely positive (e.g. positive effects on annotation quantity; Gutiérrez Páez et al. 2021; Mekler et al. 2017), yet there are little investigations into the effects of progress indicators specifically. Instantiations in this stream feature several gamification elements that realise sophisticated progress systems at different levels of abstraction (Alaghbari et al. 2020; Feyisetan et al. 2015; Jin et al. 2020): badges can for example indicate progress on individual annotation subtasks, and a leaderboard can indicate global progress (Tzerefos et al. 2022). Ultimately, showing progress appears to be a subtle, yet effective solution. Despite the lack of abstract artifacts, the vast number of instantiations in this stream provide ample inspiration for designing progress systems through gamification elements. Hence, we consider this a high maturity solution stream.

***Solution Stream 9: To support continuous engagement, adapt to individual differences in annotators.*** Gamified annotation systems may be used by a heterogeneous group of annotators with varying preferences and levels of expertise, which may require different gamification elements to continuously engage them (Dumitrache et al. 2013; Hantke et al. 2019). Thus, one solution is to adapt gamification elements to individual differences in annotators, either by tailoring them to individual annotators, or by allowing annotators to customise them (Lessel et al. 2022; Sevastjanova

et al. 2021). For this solution stream, a valuable design resource are user type models such as the gamification user type Hexad (Altmeyer et al. 2022; Y. M. Chen 2019), which even provides explicit suggestions on what gamification elements to include for which user type (Tondello et al. 2016). The general premise is that adapting gamification elements to individual preferences can improve continuous engagement and improve annotation quality and quantity (Altmeyer et al. 2022; Y. M. Chen 2019; Dumitrache et al. 2013). However, descriptive knowledge from several experiment studies in the manipulation (e.g. Altmeyer et al. 2022; Feyisetan et al. 2015) and elucidation quadrants (e.g. Y. M. Chen 2019; Hantke et al. 2019) does not fully support this premise. For example, when comparing the effects of tailored and counter-tailored versions of a gamified annotation system, Altmeyer et al. (2022) found that tailoring significantly increased user enjoyment and affective experiences yet had no effect on annotation quality or quantity. Other studies found no (Lessel et al. 2019; Lessel et al. 2022) or even negative (Schubhan et al. 2020) effects of adaptivity on annotation quality, while effects on annotation quantity were positive (Lessel et al. 2019; Schubhan et al. 2020) or neutral (Lessel et al. 2022). Remarkably, in one study (Lessel et al. 2022), annotators themselves sometimes described the adaptivity as unnecessary. Overall, adapting gamification elements is an intuitive solution and user type models like Hexad provide solid design guidance. However, even when using existing design guidance to create instantiations, the effectiveness of this solution remains questionable. Therefore, we consider this a low maturity solution stream.

### 5.4. Goal 4: shape desirable annotation behaviours directly

As a part of creating meaningful engagement, gamified annotation systems should not only indirectly support instrumental outcomes through relevant experiential outcomes, but also support instrumental outcomes directly (Liu, Santhnam, and Webster 2017). In gamified annotation systems, instrumental outcomes are principally annotation-related (e.g. annotation quality); thus, we subsume them as the goal to shape desirable annotation behaviours directly.

*Solution Stream 10: To shape desirable annotation behaviours directly, reinforce (punish) annotation behaviours that contribute to increased (decreased) annotation quality.* The first solution to this goal is to reinforce behaviours that contribute to annotation quality and correspondingly punish behaviours that contribute to decreased annotation quality. In the reviewed literature, the by far most prevalent gamification approach to do so is to use point systems, where points are awarded for correct annotations (e.g. Hantke et al. 2015), and deducted for wrong annotations (e.g. Huang and Hamarneh 2017; Jauer, Spicher, and Deserno 2021). Such feedback can form a clear connection between annotators' effort and performance (Mekler et al. 2017) and provide feelings of accomplishment (Feyisetan et al. 2015). When it is difficult to define what a correct annotation is, other behaviours may instead be reinforced: For example, annotators can be 'upgraded to a shiny new avatar' (Feyisetan et al. 2015, 336) when continuing an annotation session instead of exiting. This stream features a large range of instantiations from studies in the construction quadrant, whose evaluations generally suggest positive effects of reinforcement and punishment (e.g. Feyisetan et al. 2015; Walter, Kölle, and Collmar 2022), yet also raise some concerns regarding punishment features. Punishment can be an effective way to dissuade cheating behaviour (Mairittha and Inoue 2018; Mekler et al. 2013b), but can also discourage annotators and lead to them disengaging with an annotation system altogether (Eryiğit, Şentaş, and Monti 2022), especially when just starting out an annotation task (Fort et al. 2020). Overall, reinforcement is a more consistent solution, while punishment features are riskier yet potentially valuable against cheating behaviour. For this solution stream to mature, we require additional insights on how to design gamification elements in a way that avoids the potential negative effects from punishment. Ultimately, we consider this a medium maturity solution stream.

***Solution Stream 11: To shape desirable annotation behaviours directly, encourage timely annotation.*** While annotation quality is often considered the main instrumental outcome of gamified annotation systems (e.g. Warsinsky et al. 2022), it is also desirable for annotators to finish their annotations in a timely manner, especially when annotators are only available for a short amount of time (e.g. medical experts; Dumitrache et al. 2013; Warsinsky et al. 2022). To this end, it is an intuitive approach to couple gamification elements with annotation quantity, for example by providing points (Lessel et al. 2019) or badges for adding annotations (Chortaras et al. 2018). There are several empirical results that gamification elements can increase annotation quantity by supporting annotators' engagement to provide more annotations (Anagnostou, Mollas, and Tsoumakas 2018), thus reducing the time for individual annotations (Mekler et al. 2013b; Na and Han 2023) and for the annotation task overall (Huang and Hamarneh 2017). While some studies present gamified annotation system instantiations that support both annotation quantity and

annotation quality at the same time (Mairittha and Inoue 2018), higher annotation quantity may however be associated with lower annotation quality (Mekler et al. 2017). Yet, at present there is little guidance on how to design gamification approaches that support both annotation quality and timely annotation at the same time. Therefore, we overall consider this solution stream to be of medium maturity.

### 5.5. Goal 5: save resources

Annotating datasets requires a substantial amount of resources (e.g. money to pay annotators), which are often scarce in sectors where annotation is important (e.g. cultural heritage preservation; Ivanjko 2019; Toumanidis et al. 2019). Therefore, an important goal of gamifying an annotation system is to save resources. We found two associated solutions.

***Solution Stream 12: To save resources, replace monetary rewards with gamified rewards.*** Reimbursing annotators often requires a lot of financial resources, especially when needing many annotators or expert annotators (Dumitrache et al. 2013; Warsinsky et al. 2022). Hence, to save money, a popular solution is to replace monetary rewards with gamified rewards (e.g. Feyisetan et al. 2015; Friðriksdóttir and Einarsson 2022; Ivanjko 2019; Öhman and Kajava 2018). Many gamification elements such as points or badges are easy to implement as one-to-one substitutes for monetary rewards in gamified annotation systems (e.g. Hantke et al. 2015). Jin et al. (2020) present a model of monetary payment mechanisms for annotators (e.g. team contests), which are also mostly transferable to gamified rewards and can thus inform the design of gamified annotation systems. Generally, gamified rewards are associated with promising results (e.g. having a good cost–benefit ratio; Jauer, Spicher, and Deserno 2021). Yet, when directly comparing the effectiveness of monetary rewards and gamified rewards as furtherance incentives (i.e. incentives offered to annotators when they seek to quit the annotation task to induce them to continue), Feyisetan et al. (2015) found that monetary rewards beat gamified rewards convincingly. It currently remains unclear how to design gamified rewards so they can serve as a one-to-one replacement for monetary rewards. Thus, while gamified rewards are easy to implement and successfully applied by many existing instantiations, we consider this a medium maturity solution stream.

***Solution Stream 13: To save resources, enable crowdsourcing instead of relying on expert annotators.*** Some annotation tasks require expert annotators to perform them adequately, which are costly to acquire (e.g. medical experts for medical images; Dumitrache et al. 2013; Warsinsky et al. 2022). Against this backdrop, many studies in the reviewed literature suggest that gamification enables the use of crowdsourcing as an alternative to expert annotators (e.g. Duhaime et al. 2023; Hantke et al. 2015; Ivanjko 2019). Gamification 'provides the possibility of involving large numbers of users' (Rosani, Boato, and De Natale 2015, 1363) and aligns well with crowdsourcing annotation practices (e.g. obtaining multiple annotations per data instance; Jauer, Spicher, and Deserno 2021). We found several abstract artifacts related to this solution, including a framework to guide composition of crowd-truth gathering workflows (Dumitrache et al. 2013), a model to guide the creation of gamified crowdsourcing apps (Friðriksdóttir and Einarsson 2022) and a framework for acquisition of sensor data annotation from crowdsourcing using gamification (L'Heureux et al. 2017). Despite most of the reviewed studies featuring a crowd as annotators and the existence of three abstract artifacts in this stream, we lack insights into the mechanisms of how gamification specifically enables crowdsourcing. Most studies in this stream focus on aspects related to replacing expert annotators with a crowd (e.g. Dumitrache et al. 2013), and do not go in-depth into their gamification design. Overall, the existing abstract artifacts along with the multitude of gamified annotation system instantiations applied to crowdsourced annotation tasks provide ample design inspiration and appear to work well. However, for this solution stream to mature, we require more detailed insights into how gamification specifically enables crowdsourcing. Therefore, we consider this stream to be medium maturity.

## 6. Discussion

### 6.1. Towards more mature design knowledge on gamified annotation systems

This study aimed to capture the current state and maturity of design knowledge on gamified annotation systems and pave the way to derive more complete and mature design knowledge on gamified annotation systems. We summarise our key findings in Table 11.

Interpreting our results on artifact types and design research activities at face value suggests that design knowledge on gamified annotation systems is immature: much of the research still centres around the creation and evaluation of new instantiations in specific individual and situational contexts, while abstract design artifacts are few and there is little observation of existing artifacts. However, we noticed a

**Table 11.** Summary of key findings.

| Perspective | Previous research gaps | Our approach & key findings |
|---|---|---|
| Artifact types & design research activities | Research on gamified annotation system artifacts is scattered across annotation domains, featuring different annotation purposes, annotators, and gamification elements. We lack an overview of relevant design artifacts and research activities. | • Using the framework of DSR artifact types (Gregor and Hevner 2013) and the DRA framework (Maedche, Gregor, and Parsons 2021), we describe the landscape of artifacts and design research activities in research on gamified annotation systems.<br>• Current research on gamified annotation systems centres around the creation and evaluation of new instantiations in specific contexts.<br>• Abstract artifacts lack and existing artifacts are rarely re-used, which inhibits build-up of cumulative design knowledge for gamified annotation systems as a whole. |
| Gamification solution streams | A number of gamified annotation system instantiations are scattered across literature. We lack insights into relevant gamification solutions and whether and if so, how existing design knowledge ingrained into instantiations is mature enough to apply to gamified annotation systems as a whole. | • We derived and gauged the maturity of 13 solution streams, which provide a comprehensive overview of the gamified annotation system solution space.<br>• Our investigation of solution stream maturities reveals that the main inhibitors for mature design knowledge are that we lack (1) insights into the effectiveness of some solutions, (2) design guidance on how to avoid negative effects of gamification, and (3) abstract design artifacts that provide actionable design prescriptions.<br>• For research to create more mature design knowledge on gamified annotation systems, we recommend to (1) move beyond the creation of new instantiations as the major research activity and (2) focus on the creation of prescriptive knowledge. |

comparatively large amount of descriptive knowledge in the reviewed literature, mostly stemming from studies that did experiments with instantiations in the manipulation quadrant (e.g. studies on leaderboards; Lessel et al. 2022; Na and Han 2023). Much of this knowledge could in principle be applied to the design of gamified annotation systems in general, yet it remains unclear whether this is appropriate. Therefore, we find ourselves in a position where we have lots of scattered 'tentative' knowledge which could form the base of new knowledge strands on gamified annotation systems, which however have not been individually matured or cumulated.

Looking at the derived solution streams reveals large discrepancies in terms of maturity. While some solution streams may be considered mature (e.g. *meaningful framing to provide hedonic value*), we identified three overarching inhibitors to solution stream maturity, which also relate to gamification research in general. First, for some solutions, it was unclear whether they are important and effective in addressing their associated goal. These solutions reflect ongoing or fairly recent discourses in gamification research: For example, the solution *adapting to individual differences in annotators* reflects the ongoing call in gamification research to move from one-size-fits-all gamification to approaches adapted to individual users (Chan et al. 2024; Sezgin and Yüzer 2022). This suggests that although gamification research these days is mostly centred around understanding the inner workings of gamification (Koivisto and Hamari 2019; Nacke and Deterding 2017), for some solutions in gamified annotation systems, it may be worthwhile to take a step back and return to the question of 'does it work?' (Hamari, Koivisto, and Sarsa 2014). The second inhibitor we found was a lack of design guidance on avoiding negative side effects of gamification. Negative side effects or risks of gamification are an important discourse in gamification research (Toda, Valle, and Isotani 2017), usually associated with ethical considerations (Tseng et al. 2023). Research on human intelligence tasks (like annotation; Berg et al. 2018) has also voiced ethical concerns that gamification could contribute to exploiting workforces by motivating people to work more for less pay and encouraging unpaid labour (Ferrer-Conill 2018; Kim 2018). To this end, our reviewed literature often acknowledges risks of gamification (e.g. unwanted social comparison; Jin et al. (2020) or cheating; Mairittha and Inoue (2018)) and that there are important ethical implications when gamifying annotation tasks (Friðriksdóttir and Einarsson 2022; Gurav, Parkar, and Kharwar 2020). However, findings here are not conclusive: while most studies appear to implicitly share the aforementioned concerns about gamification exploiting workforces, some propose that gamification can even resolve some of these ethical considerations (Gurav, Parkar, and Kharwar 2020). Overall, this highlights the complexity of ethical considerations around gamification and emphasises the

importance of understanding 'the dark side of gamification' (Toda, Valle, and Isotani 2017) to support successful gamification design. Lastly, we found several solution streams to be inhibited by the lack of actionable design guidance for specific solutions. While many streams included studies that provided insightful descriptive knowledge into the inner workings of specific solutions, this was rarely supported by actionable design prescriptions that could help to implement these solutions into real-world systems. In some solution streams, this was compensated by a plethora of existing instantiations which could serve as inspiration on a solution (e.g. *showing progress*) or by gamification elements being very intuitive to implement (e.g. *providing variety*). Overall, this is reflective of the overall state of gamification research, where historically research efforts still focus on creating descriptive knowledge to gain a deeper understanding of the inner workings of gamification (Koivisto and Hamari 2019; Nacke and Deterding 2017). While such deeper insights into the inner workings of gamification are valuable, our findings support the notion that for a more complete design knowledge base, gamification research should complement the vast base of existing descriptive knowledge with equally mature prescriptive knowledge to inform artifact design.

Based on our findings, we formulate two explicit recommendations to create more mature design knowledge for gamified annotation systems. First, we recommend that research on gamified annotation systems should aim to transcend the creation of new instantiations as the major research activity. This includes shifting from creating artifacts in individual and situational contexts to deploying existing artifacts in different context, while also striving to replace instantiations as the go-to artifact of investigation in favour of more abstract artifacts like models or design principles (Gregor and Hevner 2013). Doing so would allow researchers to cumulate the valuable design knowledge ingrained in existing gamified annotation system instantiations to more clearly delineate boundaries (i.e. does the knowledge hold outside of specific application domains?; Gregor and Hevner 2013). Second, we recommend research on gamified annotation systems to focus on the creation of prescriptive knowledge, with a particular emphasis on actability of design prescriptions (Iivari, Rotvit Perlt Hansen, and Haj-Bolouri 2021). While we have a good understanding of several solutions (i.e. descriptive knowledge), we lack actionable design prescriptions to turn this understanding into solutions in real-world systems. More mature prescriptive knowledge would even the scales and support the exchange between descriptive and prescriptive knowledge bases that is vital in the creation of mature design knowledge bases (Gregor and Hevner 2013; Vom Brocke, Hevner, and Maedche 2020).

### 6.2. Contributions to research and practice

Our study makes several contributions to research and practice. Principally, our derived solution streams comprehensively capture the solution space associated with gamified annotation systems. By clearly marking out these solution streams, we move beyond fragmented insights on gamified annotation systems to allow researchers and practitioners to effectively leverage extant design knowledge. For research, we highlight those solution streams that still lack maturity and provide insights into the reasons for this. Thereby, we effectively showcase how the maturity of individual solution streams can be strengthened and provide a strong foundation for the creation of more mature design knowledge on gamified annotation systems. Compared to previous literature reviews in related domains like gamified crowdsourcing (Morschheuser et al. 2017) or gamified learning technologies (Carmichael, MacEachen, and Archibald 2022), our findings are specific to gamified annotation systems as a distinct system class. Thereby, our findings are not limited to specific annotation domains or annotators but capture the state of design knowledge on gamified annotation systems as a whole. By going beyond an overview and instead giving insights into the maturity of individual solution streams, we highlight effective ways to build more sophisticated frameworks or theory on gamified annotation systems. To this end, the explicit recommendations from the previous section also provide abstract guiding principles for effective research designs on gamified annotation systems as a whole.

Our findings also offer a valuable perspective on the maturity and scholarly progression of gamification research. Researchers have recently identified a 'second wave' of gamification research which is characterised by substantial increases in methodological rigour and the creation of full-fledged theories (Koivisto and Hamari 2019; Nacke and Deterding 2017). However, most of this research focuses on questions about *how* gamification works (Nacke and Deterding 2017); therefore, knowledge advancements constitute mostly descriptive knowledge. In contrast, our study in the context of gamified annotation systems reveals a maturity gap of prescriptive knowledge. Much of the existing prescriptive knowledge remains tied to isolated instantiations, with limited efforts to create more abstract artifacts. To foster a comparable second wave of gamification research for prescriptive knowledge, scholars should seek to transcend instantiations and work towards

developing more abstract design artifacts such as design principles or design theory. Such artifacts could provide actionable prescriptions across a range of gamification contexts (Gregor and Hevner 2013). To this end, we think that our recommendations from the previous section can also help to build a more mature prescriptive knowledge base for gamification design in general. By doing so, our study contributes to the broader efforts of elevating gamification design research to a new level of theoretical and practical maturity.

For practitioners seeking to create gamified annotation system instantiations, our findings break down the gamified annotation system design space into manageable solution streams, which outline solutions that can be individually integrated into gamified annotation system designs based on design goals. By outlining the contributors and inhibitors to the maturity of individual solution streams, we allow practitioners to see which solution streams are mature enough to be effectively employed in practice and which solutions warrant some care when implementing them in practice. Moreover, our findings pave the way for actionable prescriptions on creating gamified annotation system instantiations in practice. Thereby, we help to establish gamification as a promising and valuable tool to support manual AI training data annotation tasks, and ultimately contribute to the development of more successful gamified annotation systems in practice.

### 6.3. Limitations & future research

We acknowledge several limitations of our study, which also open avenues for future research. First, we only engaged with existing gamified annotation system artifacts in literature. While we think scientific literature hosts the main source of design knowledge on this topic, we acknowledge that important discourses surrounding AI training data annotation tools and artifacts are happening in other sources than scientific papers, such as blog posts (e.g. Alvi 2024) or organisation websites (e.g. SuperAnnotate 2023). Therefore, future research may find it useful to include other knowledge sources aside from scientific literature. Another limitation of our study is our lack of research context. To provide a comprehensive overview of existing design knowledge, we chose to investigate gamified annotation systems as a whole, instead of in specific annotation contexts (e.g. medical image annotation). As such, our results may not portray the state of knowledge in specific AI training data annotation contexts, as we may have missed some resources in these specific contexts (e.g. annotation guidelines specifically for medical image annotation; Rädsch et al. 2023). While we think it is important to abstract design knowledge away from individual contexts, future research may find it useful to tune into specific contexts to investigate specific gamification design elements.

Furthermore, for the present study, we deliberately focused on deriving solution streams for the system class of gamified annotation systems. Accordingly, our findings are also limited to the scope of gamified annotation systems. However, we did identify some resemblances between the discourses in the solution streams and ongoing discourses in gamification research (e.g. on negative side effects of gamification; Tseng et al. 2023), which indicate that some of our findings may be generalisable to gamified IS as a whole. While this study sought to contribute specifically to gamified annotation systems research, we therefore invite future research to use our findings to build higher-level frameworks that could generalise to other types of gamified IS, possibly by using established theoretical frameworks in gamification research (e.g. self-determination theory; Ryan and Deci 2020).

Lastly, while we adopted a DSR perspective in this paper and think that our results can provide valuable starting points in designing gamified annotation systems, our study does not provide actionable design prescriptions on how to create such artifacts. We think it is vital that future research takes up our results and creates actionable prescriptions for prospective designers of gamified annotation systems, for example in the form of design principles (Iivari, Rotvit Perlt Hansen, and Haj-Bolouri 2021).

## 7. Conclusion

Gamification is a valuable means to motivate and engage annotators to support the creation of high-quality AI training data. Our study highlights that contemporary research on gamified annotation systems focuses mostly on the creation and evaluation of instantiations, while mature design artifacts lack. While these instantiations have laid a valuable cornerstone, a crucial next step in building a mature design knowledge base for gamified annotation systems is to now heighten the abstraction level and build artifacts that can inform the design of gamified annotation systems as a class of systems. It remains to be investigated how well the knowledge ingrained in existing instantiations translates to the design of gamified annotation systems in general.

## Acknowledgements

The present contribution is supported by the Helmholtz Association under the joint research school 'HIDSS4Health

– Helmholtz Information and Data Science School for Health'.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## References


Alaghbari, S., A. Mitschick, G. Blichmann, M. Voigt, and R. Dachselt. 2020. "Achiever or Explorer? Gamifying the Creation Process of Training Data for Machine Learning." *Proceedings of Mensch und Computer 2020*, 173–181. Magdeburg, Sep. 06–09.

Alaghbari, S., A. Mitschick, G. Blichmann, M. Voigt, and R. Dachselt. 2021. "A User-Centered Approach to Gamify the Manual Creation of Training Data for Machine Learning." *i-com* 20 (1): 33–48. https://doi.org/10.1515/icom-2020-0030

Altmeyer, M., B. Zenuni, H. Spelt, T. Jegen, P. Lessel, and A. Krüger. 2022. "Do Hexad User Types Matter? Effects of (non-) Personalized Gamification on Task Performance and User Experience in an Image Tagging Task." *Proceedings of the ACM on Human-Computer Interaction* 6: 1–27. https://doi.org/10.1145/3549491

Alvi, F. 2024. *Data Annotation – A Beginner's Guide*. Accessed 1 August 2025. https://opencv.org/blog/data-annotation/.

Anagnostou, A., I. Mollas, and G. Tsoumakas. 2018. "Hatebusters: A Web Application for Actively Reporting YouTube Hate Speech." *27th International Joint Conference on Artificial Intelligence*, 5796–5798. Stockholm, Sweden, Jul. 13–19.

Balducci, F., and P. Buono. 2018. "Building a Qualified Annotation Dataset for Skin Lesion Analysis Trough Gamification." *Proceedings of the 2018 International Conference on Advanced Visual Interfaces*, 1–5. Castiglione della Pescaia Grosseto, Italy, May 29–Jun 01.

Batini, C., C. Cappiello, C. Francalanci, and A. Maurino. 2009. "Methodologies for Data Quality Assessment and Improvement." *ACM Computing Surveys (CSUR)* 41 (3): 1–52. https://doi.org/10.1145/1541880.1541883

Berg, J., M. Furrer, E. Harmon, U. Rani, and M. S. Silberman. 2018. *Digital Labour Platforms and the Future of Work: Towards Decent Work in the Online World*. Geneva, Switzerland: International Labour Office.

Bos, J., and M. Nissim. 2015. "Uncovering Noun-Noun Compound Relations by Gamification." *20th Nordic Conference of Computational Linguistics*, 251–255, Vilnius, Lithuania, May.

Callaghan, W., J. Goh, M. Mohareb, A. Lim, and E. Law. 2018. "Mechanicalheart: A Human-Machine Framework for the Classification of Phonocardiograms." *Proceedings of the ACM on Human-Computer Interaction* 2 (CSCW): 1–17. https://doi.org/10.1145/3274297

Cao, H.-A., T. K. Wijaya, K. Aberer, and N. Nunes. 2015. "A Collaborative Framework for Annotating Energy Datasets." *2015 IEEE International Conference on Big Data*, 2716–2725. Santa Clara, CA, USA, Oct. 29–Nov. 01.

Carmichael, D., C. MacEachen, and J. Archibald. 2022. "Gamification in a Learning Resource for the Study of Human Computer Interaction." *Intelligent Computing: Proceedings of the 2021 Computing Conference*, 697–717, Virtual Conference, Jul. 15–16.

Chamberlain, J. 2014. "The Annotation-Validation (AV) Model." *First International Workshop on Gamification for Information Retrieval*, 12–16. Amsterdam, Netherlands, Apr. 13.

Chan, G., A. Arya, R. Orji, Z. Zhao, and A. Whitehead. 2024. "Increasing Motivation in Social Exercise Games: Personalising Gamification Elements to Player Type." *Behaviour & Information Technology* 43 (11): 2608–2638. https://doi.org/10.1080/0144929X.2023.2255293

Chandler, D., and A. Kapelner. 2013. "Breaking Monotony with Meaning: Motivation in Crowdsourcing Markets." *Journal of Economic Behavior & Organization* 90:123–133. https://doi.org/10.1016/j.jebo.2013.03.003

Chen, C.-M., M.-C. Li, and T.-C. Chen. 2020. "A web-Based Collaborative Reading Annotation System with Gamification Mechanisms to Improve Reading Performance." *Computers & Education* 144:103697. https://doi.org/10.1016/j.compedu.2019.103697

Chen, Y.-M. 2019. "Motivational Design in Translation Crowdsourcing: A Gamification Approach to Facebook Community Translation." *Compilation & Translation Review* 12 (1): 141–176.

Chortaras, A., A. Christaki, N. Drosopoulos, E. Kaldeli, M. Ralli, A. Sofou, A. Stabenau, G. Stamou, and V. Tzouvaras. 2018. "WITH: Human-Computer Collaboration for Data Annotation and Enrichment." *Companion Proceedings of the The Web Conference 2018*, 1117–1125, Lyon, France, Apr. 23–27.

Cohen, J. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20 (1): 37–46. https://doi.org/10.1177/001316446002000104

Deterding, S., D. Dixon, R. Khaled, and L. Nacke. 2011. "From Game Design Elements to Gamefulness: Defining Gamification." *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, 9–15, Tampere, Finland, Sep. 28–30.

Duhaime, E. P., M. Jin, T. Moulton, J. Weber, N. R. Kurtansky, A. Halpern, and V. Rotemberg. 2023, Apr. 18–21. "Nonexpert Crowds Outperform Expert Individuals in Diagnostic Accuracy on a Skin Lesion Diagnosis Task." *IEEE 20th International Symposium on Biomedical Imaging, Cartagena de Indias, Colombia*.

Dumitrache, A., L. Aroyo, C. Welty, R.-J. Sips, and A. Levas. 2013, Oct. 21. "Dr. Detective: Combining Gamification Techniques and Crowdsourcing to Create a Gold Standard for the Medical Domain." *1st International Workshop on Crowdsourcing the Semantic Web, Sydney, Australia*.

Ertan, K., and S. A. Kocadere. 2022. "Gamification Design to Increase Motivation in Online Learning Environments: A Systematic Review." *Journal of Learning and Teaching in Digital Age* 7 (2): 151–159. https://doi.org/10.53850/joltida.1020044

Eryiğit, G., A. Şentaş, and J. Monti. 2022. "Gamified Crowdsourcing for Idiom Corpora Construction." *Natural Language Engineering* 29 (4): 909–941. https://doi.org/10.1017/S1351324921000401

Ferrer-Conill, R. 2018. "Playbour and the Gamification of Work: Empowerment, Exploitation and fun as Labour

Dynamics." In *Technologies of Labour and the Politics of Contradiction*, edited by Bilić Paško, Primorac Jaka, and Valtýsson Bjarki, 193–210. Cham, Switzerland: Springer.

Feyisetan, O., E. Simperl, M. Van Kleek, and N. Shadbolt. 2015. "Improving Paid Microtasks through Gamification and Adaptive Furtherance Incentives," *Proceedings of the 24th International Conference on World Wide Web*, 333–343, Florence, Italy, May 18.

Fort, K., B. Guillaume, Y.-A. Pilatte, M. Constant, and N. Lefèbvre. 2020. "Rigor Mortis: Annotating MWEs with a Gamified Platform." *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4395–4401, Marseille, France, May 13–15.

Friðriksdóttir, S. R., and H. Einarsson. 2022. "Fictionary-Based Games for Language Resource Creation." *Proceedings of the 2nd Workshop on Novel Incentives in Data Collection from People: Models, Implementations, Challenges and Results within LREC 2022*, 25–31, Marseille, France, Jun.

Gregor, S., and A. R. Hevner. 2013. "Positioning and Presenting Design Science Research for Maximum Impact." *MIS Quarterly* 37 (2): 337–355. https://doi.org/10.25300/MISQ/2013/37.2.01

Gregor, S., and D. Jones. 2007. "The Anatomy of a Design Theory." *Journal of the Association for Information Systems* 8 (5): 312–335. https://doi.org/10.17705/1jais.00129

Gudivada, V., A. Apon, and J. Ding. 2017. "Data Quality Considerations for big Data and Machine Learning: Going beyond Data Cleaning and Transformations." *International Journal on Advances in Software* 10 (1): 1–20.

Guillot, L., Q. Bragard, R. Smith, and A. Ventresque. 2016. "Towards a Gamified System to Improve Translation for Online Meetings." *Proceedings of the Third International Workshop on Gamification for Information Retrieval*, 21–28, Pisa, Italy, Jul. 21.

Gurav, V., M. Parkar, and P. Kharwar. 2020. "Accessible and Ethical Data Annotation with the Application of Gamification." *International Conference on Recent Developments in Science, Engineering and Technology*, 68–78, New York, NY, USA, Jun. 01–02.

Gutiérrez Páez, N. F., J. S. Gómez-Cañón, L. Porcaro, P. Santos, D. Hernández-Leo, and E. Gómez. 2021. "Emotion Annotation of Music: A Citizen Science Approach." *International Conference on Collaboration Technologies and Social Computing*, 51–66, Virtual Conference, Aug. 31–Sep. 01.

Hamari, J., J. Koivisto, and H. Sarsa. 2014. "Does Gamification Work? A Literature Review of Empirical Studies on Gamification." *47th Hawaii International Conference on System Sciences*, 3025–3034, Waikoloa, HI, USA, Jan. 06–09.

Hantke, S., A. Abstreiter, N. Cummins, and B. Schuller. 2018. "Trustability-Based Dynamic Active Learning for Crowdsourced Labelling of Emotional Audio Data." *IEEE Access* 6:42142–42155. https://doi.org/10.1109/ACCESS.2018.2858931

Hantke, S., T. Appel, and B. Schuller. 2018. "The Inclusion of Gamification Solutions to Enhance User Enjoyment on Crowdsourcing Platforms." *1st Asian Conference on Affective Computing and Intelligent Interaction*, 1–6, Beijing, China, May 20–22.

Hantke, S., F. Eyben, T. Appel, and B. Schuller. 2015. "iHEARu-PLAY: Introducing a Game for Crowdsourced Data Collection for Affective Computing." *International Conference on Affective Computing and Intelligent Interaction*, 891–897, Xi'an, China, Sep. 21–24.

Hantke, S., T. Olenyi, C. Hausner, T. Appel, and B. Schuller. 2019. "Large-scale Data Collection and Analysis via a Gamified Intelligent Crowdsourcing Platform." *International Journal of Automation and Computing* 16 (4): 427–436. https://doi.org/10.1007/s11633-019-1180-0

Hantke, S., Z. Zhang, and B. Schuller. 2017. "Towards Intelligent Crowdsourcing for Audio Data Annotation: Integrating Active Learning in the Real World." *18th Annual Conference of the International Speech Communication Association*, 3951–3955, Stockholm, Sweden, Aug. 20–24.

Hevner, A. R., S. T. March, J. Park, and S. Ram. 2004. "Design Science in Information Systems Research." *MIS Quarterly* 28 (1): 75–105. https://doi.org/10.2307/25148625

Huang, M., and G. Hamarneh. 2017. "SwifTree: Interactive Extraction of 3D Trees Supporting Gaming and Crowdsourcing." *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 6th Joint International Workshops, CVII-STENT 2017 and Second International Workshop, LABELS 2017, Held in Conjunction with MICCAI 2017*, 116–125, Québec City, QC, Canada, Sep. 10–14.

Iivari, J., M. Rotvit Perlt Hansen, and A. Haj-Bolouri. 2021. "A Proposal for Minimum Reusability Evaluation of Design Principles." *European Journal of Information Systems* 30 (3): 286–303. https://doi.org/10.1080/0960085X.2020.1793697

Ivanjko, T. 2019. "Crowdsourcing Image Descriptions Using Gamification: A Comparison Between Game-Generated Labels and Professional Descriptors." *42nd International Convention on Information and Communication Technology, Electronics and Microelectronics*, 537–541, Opatija, Croatia, May 20–24.

Jauer, M.-L., N. Spicher, and T. M. Deserno. 2021. "Gamification Concept for Acquisition of Medical Image Segmentation via Crowdsourcing." *SPIE Medical Imaging 2021: Imaging Informatics for Healthcare, Research, and Applications*, 77–85, Virtual Conference, Feb. 15–20.

Jin, Y., M. Carman, Y. Zhu, and Y. Xiang. 2020. "A Technical Survey on Statistical Modelling and Design Methods for Crowdsourcing Quality Control." *Artificial Intelligence* 287:103351. https://doi.org/10.1016/j.artint.2020.103351

Khalaji, E., S. Eraslan, Y. Yesilada, and V. Yaneva. 2023. "Effects of Data Preprocessing on Detecting Autism in Adults Using web-Based eye-Tracking Data." *Behaviour & Information Technology* 42 (14): 2476–2484. https://doi.org/10.1080/0144929X.2022.2127376

Kicikoglu, D., R. Bartle, J. Chamberlain, and M. Poesio. 2019. "Wormingo: A 'True Gamification' Approach to Anaphoric Annotation." *Proceedings of the 14th International Conference on the Foundations of Digital Games*, 1–7, San Luis Obispo, CA, USA, Aug. 26–30.

Kim, T. W. 2018. "Gamification of Labor and the Charge of Exploitation." *Journal of Business Ethics* 152 (1): 27–39. https://doi.org/10.1007/s10551-016-3304-6

Koivisto, J., and J. Hamari. 2019. "The Rise of Motivational Information Systems: A Review of Gamification Research." *International Journal of Information Management* 45:191–210. https://doi.org/10.1016/j.ijinfomgt.2018.10.013

Lessel, P., M. Altmeyer, L. V. Schmeer, and A. Krüger. 2019. "Enable or Disable Gamification? Analyzing the Impact of Choice in a Gamified Image Tagging Task." *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12, Glasgow, Scotland UK, May 4–9.

Lessel, P., M. Altmeyer, M. Schubhan, D. Gudea, and A. Krüger. 2022. "Absolute or Relative? Exploring the Choice Between Leaderboard Types in an Image Tagging Task." *Extended Abstracts of the Annual Symposium on Computer-Human Interaction in Play*, 63–69.

Levy, Y., and T. J. Ellis. 2006. "A Systems Approach to Conduct an Efective Literature Review in Support of Information Systems Research." Informing Science 9: 181–212. https://doi.org/10.28945/479.

L'Heureux, A., K. Grolinger, W. A. Higashino, and M. A. M. Capretz. 2017. "A Gamification Framework for Sensor Data Analytics." *2017 IEEE International Congress on Internet of Things*, 74–81, Honolulu, HI, USA, Jun. 25–30.

Liu, D., R. Santhanam, and J. Webster. 2017. "Toward Meaningful Engagement: A Framework for Design and Research of Gamified Information Systems." *MIS Quarterly* 41 (4): 1011–1034. https://doi.org/10.25300/MISQ/2017/41.4.01

Maedche, A., S. Gregor, S. Morana, and J. Feine. 2019. "Conceptualization of the Problem Space in Design Science Research." *14th International Conference on Design Science Research in Information Systems and Technology*, 18–31, Worcester, MA, USA, Jun. 04–06.

Maedche, A., S. Gregor, and J. Parsons. 2021. "Mapping Design Contributions in Information Systems Research: The Design Research Activity Framework." *Communications of the Association for Information Systems* 49 (1): 355–378. https://doi.org/10.17705/1CAIS.04914.

Mairittha, N., and S. Inoue. 2018. "Gamification for High-Quality Dataset in Mobile Activity Recognition." *International Conference on Mobile Computing, Applications, and Services*, 216–222, Osaka, Japan, Feb. 28–Mar. 02.

Mairittha, N., T. Mairittha, and S. Inoue. 2019. "Optimizing Activity Data Collection with Gamification Points Using Uncertainty Based Active Learning." *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, 761–767.

Mairittha, N., T. Mairittha, P. Lago, and S. Inoue. 2021. "CrowdAct: Achieving High-Quality Crowdsourced Datasets in Mobile Activity Recognition." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5 (1): 1–32. https://doi.org/10.1145/3432222

Maltzahn, C., A. Jhala, M. Mateas, and J. Whitehead. 2014. "Gamification of Private Digital Data Archive Management." *Proceedings of the First International Workshop on Gamification for Information Retrieval*, 33–37, Amsterdam, Netherlands, Apr. 13.

Megorskaya, O., V. Kukushkin, and P. Serdyukov. 2015. "On the Relation Between Assessor's Agreement and Accuracy in Gamified Relevance Assessment." *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 605–614.

Meireles, O. R., G. Rosman, M. S. Altieri, L. Carin, G. Hager, A. Madani, N. Padoy, C. M. Pugh, P. Sylla, and T. M. Ward. 2021. "SAGES Consensus Recommendations on an Annotation Framework for Surgical Video." *Surgical Endoscopy* 35 (9): 4918–4929. https://doi.org/10.1007/s00464-021-08578-9

Mekler, E., F. Brühlmann, K. Opwis, and A. N. Tuch. 2013a. "Disassembling Gamification: The Effects of Points and Meaning on User Motivation and Performance." *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, 1137–1142, Paris, France, Apr. 27–May 02.

Mekler, E., F. Brühlmann, K. Opwis, and A. N. Tuch. 2013b. "Do Points, Levels and Leaderboards Harm Intrinsic Motivation?" *Proceedings of the First International Conference on Gameful Design, Research, and Applications*, 66–73.

Mekler, E., F. Brühlmann, A. N. Tuch, and K. Opwis. 2017. "Towards Understanding the Effects of Individual Gamification Elements on Intrinsic Motivation and Performance." *Computers in Human Behavior* 71:525–534. https://doi.org/10.1016/j.chb.2015.08.048

Millour, A., and K. Fort. 2018. "Toward a Lightweight Solution for Less-Resourced Languages: Creating a POS Tagger for Alsatian Using Voluntary Crowdsourcing." *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 455–460, Miyazaki, Japan, May 07–12.

Mora, A., D. Riera, C. González, and J. Arnedo-Moreno. 2017. "Gamification: A Systematic Review of Design Frameworks." *Journal of Computing in Higher Education* 29 (3): 516–548. https://doi.org/10.1007/s12528-017-9150-4

Morschheuser, B., J. Hamari, J. Koivisto, and A. Maedche. 2017. "Gamified Crowdsourcing: Conceptualization, Literature Review, and Future Agenda." *International Journal of Human-Computer Studies* 106:26–43. https://doi.org/10.1016/j.ijhcs.2017.04.005

Morschheuser, B., L. Hassan, K. Werder, and J. Hamari. 2018. "How to Design Gamification? A Method for Engineering Gamified Software." *Information and Software Technology* 95:219–237. https://doi.org/10.1016/j.infsof.2017.10.015

Na, K., and K. Han. 2023. "How Leaderboard Positions Shape Our Motivation: The Impact of Competence Satisfaction and Competence Frustration on Motivation in a Gamified Crowdsourcing Task." *Internet Research* 33 (7): 1–18. https://doi.org/10.1108/INTR-12-2021-0897

Nacke, L. E., and S. Deterding. 2017. "The Maturing of Gamification Research." *Computers in Human Behavior* 71:450–454. https://doi.org/10.1016/j.chb.2016.11.062

Nakamura, J., and M. Csikszentmihalyi. 2009. "Flow Theory and Research." In *Handbook of Positive Psychology*, edited by C. R. Snyder and Shane J. Lopez, 195–206. New York: Oxford University Press.

Neves, M., and U. Leser. 2014. "A Survey on Annotation Tools for the Biomedical Literature." *Briefings in Bioinformatics* 15 (2): 327–340. https://doi.org/10.1093/bib/bbs084

Neves, M., and J. Ševa. 2021. "An Extensive Review of Tools for Manual Annotation of Documents." *Briefings in Bioinformatics* 22 (1): 146–163. https://doi.org/10.1093/bib/bbz130

Öhman, E., and K. Kajava. 2018. "Sentimentator: Gamifying Fine-Grained Sentiment Annotation." *Digital Humanities in the Nordic Countries*, 98–110, Helsinki, Finland, Mar. 07.

Öhman, E., K. Kajava, J. Tiedemann, and T. Honkela. 2018. "Creating a Dataset for Multilingual Fine-Grained Emotion-Detection Using Gamification-Based Annotation." *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 24–30, Brussels, Belgium, Oct.

Page, M. J., J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, and S. E. Brennan. 2021. "The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews." *BMJ* 372:n71. https://doi.org/10.1136/bmj.n71

Palmquist, A. 2024. "Stakeholders' Design Preferences for Instructional Gamification." *Behaviour & Information Technology* 43 (12): 2717–2749. https://doi.org/10.1080/0144929X.2023.2255905

Plappert, M., C. Mandery, and T. Asfour. 2016. "The KIT Motion-Language Dataset." *Big Data* 4 (4): 236–252. https://doi.org/10.1089/big.2016.0028

Pustejovsky, J., and A. Stubbs. 2012. *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications*. Sebastopol, CA: O'Reilly Media, Inc.

Rädsch, T., A. Reinke, V. Weru, M. D. Tizabi, N. Schreck, A. E. Kavur, B. Pekdemir, T. Roß, A. Kopp-Schneider, and L. Maier-Hein. 2023. "Labelling Instructions Matter in Biomedical Image Analysis." *Nature Machine Intelligence* 5 (3): 273–283. https://doi.org/10.1038/s42256-023-00625-5

Riegler, M., R. Eg, L. Calvet, M. Lux, P. Halvorsen, and C. Griwodz. 2015. "Playing Around the Eye Tracker-A Serious Game Based Dataset." *Proceedings of the 2nd International Workshop on Gamification for Information Retrieval*, 34–40, Vienna, Austria, Mar. 29.

Rosani, A., G. Boato, and F. G. B. De Natale. 2015. "EventMask: A Game-Based Framework for Event-Saliency Identification in Images." *IEEE Transactions on Multimedia* 17 (8): 1359–1371. https://doi.org/10.1109/TMM.2015.2441003

Russo, S., M. D. Besmer, F. Blumensaat, D. Bouffard, A. Disch, F. Hammes, A. Hess, M. Lürig, B. Matthews, and C. Minaudo. 2021. "The Value of Human Data Annotation for Machine Learning Based Anomaly Detection in Environmental Systems." *Water Research* 206:117695. https://doi.org/10.1016/j.watres.2021.117695

Ryan, R. M., and E. L. Deci. 2020. "Intrinsic and Extrinsic Motivation from a Self-determination Theory Perspective: Definitions, Theory, Practices, and Future Directions." *Contemporary Educational Psychology* 61:101860. https://doi.org/10.1016/j.cedpsych.2020.101860

Sager, C., C. Janiesch, and P. Zschech. 2021. "A Survey of Image Labelling for Computer Vision Applications." *Journal of Business Analytics* 4 (2): 91–110. https://doi.org/10.1080/2573234X.2021.1908861

Schubhan, M., M. Altmeyer, D. Buchheit, and P. Lessel. 2020. "Investigating User-Created Gamification in an Image Tagging Task." *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12.

Sevastjanova, R., W. Jentner, F. Sperrle, R. Kehlbeck, J. Bernard, and M. El-assady. 2021. "QuestionComb: A Gamification Approach for the Visual Explanation of Linguistic Phenomena through Interactive Labeling." *ACM Transactions on Interactive Intelligent Systems* 11 (3-4): 1–38. https://doi.org/10.1145/3429448

Sezgin, S., and T. V. Yüzer. 2022. "Analysing Adaptive Gamification Design Principles for Online Courses." *Behaviour & Information Technology* 41 (3): 485–501. https://doi.org/10.1080/0144929X.2020.1817559

Simões, B., and R. De Amicis. 2016. "Gamification as a Key Enabling Technology for Image Sensing and Content Tagging." In *Intelligent Interactive Multimedia Systems and Services 2016*, edited by Giuseppe De Pietro, Luigi Gallo, Robert J. Howlett, and C. Lakhmi, 503–513. Cham, Switzerland: Springer.

Snyder, H. 2019. "Literature Review as a Research Methodology: An Overview and Guidelines." *Journal of Business Research* 104:333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Strohmann, T., and B. Khosrawi-Rad. 2025. "The Relevance of Design Features: From Abstract Knowledge to Practical Implementation." *International Conference on Design Science Research in Information Systems and Technology*, 116–134, Montego Bay, Jamaica, Jun. 02–04.

SuperAnnotate. 2023. Text Annotation for Machine Learning. Accessed 1 August 2025. https://www.superannotate.com/blog/text-annotation-for-machine-learning.

Toda, A. M., P. H. Valle, and S. Isotani. 2017. "The Dark Side of Gamification: An Overview of Negative Effects of Gamification in Education." *International Workshop on Social, Semantic, Adaptive and Gamification Techniques and Technologies for Distance Learning*, 143–156, Maceió, Brazil, Mar. 20–24.

Tondello, G. F., R. R. Wehbe, L. Diamond, M. Busch, A. Marczewski, and L. E. Nacke. 2016. "The Gamification User Types Hexad Scale." *CHI PLAY '16: Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*, 229–243, Austin, Texas, USA, Oct. 16–19.

Toumanidis, L., E. Bocaj, P. Kasnesis, and C. Z. Patrikakis. 2019. "Supporting Cultural Heritage Preservation Through Game-Based Crowdsourcing." *International Conference on Strategic Innovative Marketing and Tourism, pp. 989–997, Aigaleo, Attika, Greece*, Oct. 17–20.

Tseng, S.-L. A., H. Sun, R. Santhanam, S. Lu, and J. B. Thatcher. 2023. "Rethinking Gamification Failure: A Model and Investigation of Gamified System Maladaptive Behaviors." *Information Systems Research* 35 (4): 1743–1765. https://doi.org/10.1287/isre.2021.0284.

Tzerefos, A., I. Kanellos, S. Chatzopoulos, T. Dalamagas, and T. Vergoulis. 2022. "SurvAnnT: Facilitating Community-Led Scientific Surveys and Annotations." *26th International Conference on Theory and Practice of Digital Libraries*, 543–547, Padua, Italy, Sep. 20–23.

Vecchio, G., S. Palazzo, D. Giordano, F. Rundo, and C. Spampinato. 2020. "MASK-RL: Multiagent Video Object Segmentation Framework through Reinforcement Learning." *IEEE Transactions on Neural Networks and Learning Systems* 31 (12): 5103–5115. https://doi.org/10.1109/TNNLS.2019.2963282

Viana, P., and J. P. Pinto. 2017. "A Collaborative Approach for Semantic Time-Based Video Annotation Using Gamification." *Human-centric Computing and Information Sciences* 7 (1): 1–21. https://doi.org/10.1186/s13673-017-0094-5

Vom Brocke, J., A. Hevner, and A. Maedche. 2020. "Introduction to Design Science Research." In *Design Science Research Cases*, edited by Jan vom Brocke, Alan Hevner, and Alexander Maedche, 1–13. Cham, Switzerland: Springer.

Wagner, M., A. Bihlmaier, H. G. Kenngott, P. Mietkowski, P. M. Scheikl, S. Bodenstedt, A. Schiepe-Tiska, J. Vetter, F. Nickel, and S. Speidel. 2021. "A Learning Robot for Cognitive Camera Control in Minimally Invasive Surgery." *Surgical Endoscopy* 35 (9): 5365–5374. https://doi.org/10.1007/s00464-021-08509-8

Walter, V., M. Kölle, and D. Collmar. 2022. "A Gamification Approach for the Improvement of Paid Crowd-Based Labelling of Geospatial Data." *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* V4-2022:113–120. https://doi.org/10.5194/isprs-annals-V-4-2022-113-2022

Wand, Y., and R. Y. Wang. 1996. "Anchoring Data Quality Dimensions in Ontological Foundations." *Communications of the ACM* 39 (11): 86–95. https://doi.org/10.1145/240455.240479

Wang, S., C. Li, R. Wang, Z. Liu, M. Wang, H. Tan, Y. Wu, X. Liu, H. Sun, and R. Yang. 2021. "Annotation-efficient Deep Learning for Automatic Medical Image Segmentation." *Nature Communications* 12 (1): 5915. https://doi.org/10.1038/s41467-021-26216-9

Warsinsky, S., M. Schmidt-Kraepelin, S. Rank, S. Thiebes, and A. Sunyaev. 2021. "Conceptual Ambiguity Surrounding Gamification and Serious Games in Health Care: Literature Review and Development of Game-Based Intervention Reporting Guidelines (GAMING)." *Journal of Medical Internet Research* 23 (9): e30390. https://doi.org/10.2196/30390

Warsinsky, S., M. Schmidt-Kraepelin, S. Thiebes, M. Wagner, and A. Sunyaev. 2022. "Gamified Expert Annotation Systems: Meta-Requirements and Tentative Design." *International Conference on Design Science Research in Information Systems and Technology*, 154–166, St. Petersburg, FL, USA, Jun 01–03.

Washington, P., H. Kalantarian, J. Kent, A. Husic, A. Kline, E. Leblanc, C. Hou, O. C. Mutlu, K. Dunlap, and Y. Penev. 2022. "Improved Digital Therapy for Developmental Pediatrics Using Domain-Specific Artificial Intelligence: Machine Learning Study." *JMIR Pediatrics and Parenting* 5 (2): e26760. https://doi.org/10.2196/26760

Webster, J., and R. T. Watson. 2002. "Analyzing the Past to Prepare for the Future: Writing a Literature Review." *MIS Quarterly* 26 (2): xiii–xxiii. https://doi.org/10.2307/249576

Yang, C., H. J. Ye, and Y. Feng. 2021. "Using Gamification Elements for Competitive Crowdsourcing: Exploring the Underlying Mechanism." *Behaviour & Information Technology* 40 (9): 837–854. https://doi.org/10.1080/0144929X.2020.1733088

## Appendix A. Overview of solution stream maturity criteria for all solution streams

We assessed the maturity of each solution based on the presence of abstract artifacts surrounding the solution, design research activities in the solution stream, the presence of justificatory knowledge, knowledge on the effectiveness of the solution and the solution's implementability. The following table provides an overview of our assessment of each of these criteria for each solution. Note that while most criteria can have a positive or negative impact on solution stream maturity, we decided to value it as neutral if there are no abstract artifacts to inform the design of a specific solution. We did so as we sought to evaluate the maturity of individual solution streams relative to the gamified annotation system design knowledge base as a whole, where abstract artifacts generally lack.

| # | *Solution stream* | Presence of abstract artifacts in domain | Design research activities | Presence of justificatory knowledge | Knowledge on effectiveness of solution | Implementability | Solution maturity |
|---|---|---|---|---|---|---|---|
| 1 | *To ensure that annotators feel compensated for their work, provide incentives that support intrinsic motivation.* | o No abstract artifacts to inform design | + Several insightful studies in manipulation and elucidation quadrant | + Self-determination theory as an asset to inform design<br>o Average detail level in design rationales | + Lots of empirical results on this solution specifically<br>− Inconclusive whether aiming to achieve intrinsic motivation in gamified annotation systems is really expedient | − Intrinsic motivation is difficult to design for<br>− Little design guidance on solution specifically | Low |
| 2 | *To ensure that annotators feel compensated for their work, provide retrospective rewards if ground truth is not clear at time of annotation.* | + Existing model to inform design of rewards through retrospective validation specifically | o Most studies in construction quadrant, few studies in other quadrants | o Average detail level in design rationales | + Effective to circumvent issues from lack of ground truth annotations<br>+ Generally positive empirical results on instantiations using this solution<br>− Little empirical results for this solution specifically | + Actionable design guidance from existing model | High |
| 3 | *To ensure that annotators perceive value from using the annotation system, provide meaningful framing of the annotation task.* | o No abstract artifacts to inform design | + Several insightful studies in manipulation quadrant | + Detailed design rationales on solution specifically | + Strong empirical results for effectiveness<br>+ Empirical results on this solution specifically | + Intuitive to implement by embedding annotation task into narrative | High |
| 4 | *To ensure that annotators perceive value from using the annotation system, hide the annotation task behind game elements.* | o No abstract artifacts to inform design | − Almost exclusively studies in the construction quadrant | + Detailed design rationales on solution specifically | + Generally positive empirical results on instantiations using this solution<br>− Little empirical results on this solution specifically | − Lack of design guidance that informs positioning on game continuum | Med |
| 5 | *To ensure that annotators perceive value from using the annotation system, create meaningful social interactions.* | o No abstract artifacts to inform design | o Most studies in construction quadrant, few studies in other quadrants | o Design rationales mostly on overall artifact; little rationales for solution specifically | + Mostly positive empirical results<br>− Several possible negative side effects identified | − Lack of design guidance on how to avoid potential negative effects associated with this solution | Med |
| 6 | *To support continuous engagement, provide variety in how to interact with the annotation system.* | + Existing model to support design for variety specifically | o Most studies in construction quadrant, few studies in other quadrants | o Design rationales mostly on overall artifact; little rationales for solution specifically | + Generally positive empirical results on instantiations using this solution<br>− Little empirical results on this solution specifically | + Intuitive to implement: almost all gamification elements introduce variety<br>+ Design guidance from existing model<br>+ Design guidance on avoiding negative side effects | High |

(*Continued*)

Continued.

| # | *Solution stream* | Presence of abstract artifacts in domain | Design research activities | Presence of justificatory knowledge | Knowledge on effectiveness of solution | Implementability | Solution maturity |
|---|---|---|---|---|---|---|---|
| 7 | *To support continuous engagement, gradually raise task difficulty.* | o No abstract artifacts to inform design | o Most studies in construction quadrant, few studies in other quadrants | + Flow theory as an asset to inform design<br>o Design rationales mostly on overall artifact; little rationales for solution specifically | + Conceptually effective<br>− Little empirical results | + Intuitive to implement through points systems<br>− Unclear how to implement solution if ground truth annotations are not present to assess difficulty of annotating specific data instances | Low |
| 8 | *To support continuous engagement, show annotators their annotation progress.* | o No abstract artifacts to inform design; | + Large range of instantiations in construction quadrant using this solution<br>o Most studies in construction quadrant, few studies in other quadrants | o Design rationales mostly on overall artifact; little rationales for solution specifically | + Generally positive empirical results on instantiations using this solution<br>− Little empirical results on this solution specifically | + Intuitive to implement through various gamification elements | High |
| 9 | *To support continuous engagement, adapt to individual differences in annotators.* | o No abstract artifacts to inform design | + Several insightful studies in the manipulation quadrant | + Several existing user type models to guide design; for example, gamification user type Hexad | + Several empirical results on solution specifically<br>- Many neutral or mixed empirical results<br>- Importance of solution questionable; | + User type models guide which gamification elements suit each user type<br>- Usually large implementation efforts required | Low |
| 10 | *To shape desirable annotation behaviours directly, reinforce (punish) annotation behaviours that contribute to increased (decreased) annotation quality.* | o No abstract artifacts to inform design | o Most studies in construction quadrant, few studies in other quadrants<br>+ Large range of instantiations in construction quadrant using this solution | + Detailed design rationales on solution specifically | + Generally positive empirical results on instantiations using this solution<br>+ Effective to provide feedback to annotators and tackle cheating behaviour<br>- Unclear how to avoid negative side effects from punishment | + Intuitive to implement through points systems | Med |
| 11 | *To shape desirable annotation behaviours directly, encourage timely annotation.* | o No abstract artifacts to inform design | o Most studies in construction quadrant, few studies in other quadrants<br>+ Large range of instantiations in construction quadrant using this solution | o Design rationales mostly on overall artifact; little rationales for solution specifically | + Several empirical results on solution specifically<br>+ Empirical results largely positive<br>- Several studies with negative effects on annotation quality | + Intuitive to implement by coupling gamification elements with annotation quantity<br>- Unclear how to avoid jeopardising annotation quality while encouraging annotation quantity | Med |
| 12 | *To save resources, replace monetary rewards with gamified rewards.* | + Existing model to inform design | o Most studies in construction quadrant, few studies in other quadrants | + Detailed design rationales on solution specifically | - Several empirical results that monetary rewards outperform gamified rewards | + Intuitive to replace monetary rewards with gamified rewards one-to-one | Med |
| 13 | *To save resources, enable crowdsourcing instead of relying on expert annotators.* | o Several existing abstract artifacts, but abstract artifacts are only tangentially related to gamification | o Most studies in construction quadrant, few studies in other quadrants<br>+ Large range of instantiations in construction quadrant using this solution | - Mostly abstract design rationales; unclear why solution works the way it does | + Several empirical results on replacing expert annotators with crowd<br>- Little empirical results related to gamification<br>- Specific mechanisms how gamification enables crowdsourcing remain unclear | + Ample design inspiration from many existing instantiations | Med |