Designing Gamification Concepts for Expert Explainable Artificial Intelligence Evaluation Tasks: A Problem Space Exploration

Philipp A. Toussaint
Karlsruhe Institute of Technology
philipp.toussaint@kit.edu

Simon Warsinsky
Karlsruhe Institute of Technology
simon.warsinsky@kit.edu

Manuel Schmidt-Kraepelin
Karlsruhe Institute of Technology
manuel.schmidt-kraepelin@kit.edu

Scott Thiebes
Karlsruhe Institute of Technology
scott.thiebes@kit.edu

Ali Sunyaev
Karlsruhe Institute of Technology
sunyaev@kit.edu

Abstract

Artificial intelligence (AI) models are often complex and require additional explanations for use in high-stakes decision-making contexts like healthcare. To this end, explainable AI (XAI) developers must evaluate their explanations with domain experts to ensure understandability. As these evaluations are tedious and repetitive, we look at gamification as a means to motivate and engage experts in XAI evaluation tasks. We explore the problem space associated with gamified expert XAI evaluation. Based on a literature review of 22 relevant studies and seven interviews with experts in XAI evaluation, we elicit knowledge about affected stakeholders, eight needs, eight goals, and seven requirements. Our results help us understand better the problems associated with expert XAI evaluation and paint a broad application potential for gamification to improve XAI expert evaluations. In doing so, we lay the foundation for the design of successful gamification concepts for expert XAI evaluation.

Keywords: Gamification, Explainable Artificial Intelligence, Evaluation, Problem Space

1. Introduction

Artificial Intelligence (AI) models are becoming ever more prominent and proficient, but also more complex and hard to understand for humans (Adadi & Berrada, 2018). To meaningfully interpret the outputs of AI models — especially deep learning models (Mulgund et al., 2022) — explanations are often required (Barredo Arrieta et al., 2020). Especially in high-stakes contexts like healthcare, where wrong AI decisions may be fatal, understanding AI models is pivotal (Butz et al., 2022; Lötsch et al., 2021). Here, explainable AI (XAI) helps to understand and justify AI decisions, as well as increase trust in and transparency of AI models (Adadi & Berrada, 2018). XAI research has produced various methods to improve AI models' explainability (Jacovi, 2023). Yet, given a specific XAI context, it is still difficult to tell what "the best" explanation would be (Zhou et al., 2021), as the value or helpfulness of an explanation is dependent on context (Butz et al., 2022). Thus, XAI explanations must be evaluated on a per-context basis (Ben David et al., 2021). Human-centered XAI evaluations can "provide direct and strong evidence of [the] success of explanations" (Zhou et al., 2021, p. 602) and help evaluate explanations regarding their explanatory power, usefulness, and understandability by users of relevant XAI systems (Zhou et al., 2021). XAI evaluation tasks often require domain experts with deep knowledge in an application area, for example, clinical expertise for disease diagnosis (Chu et al., 2020). However, motivating experts to do XAI evaluation tasks can be difficult, as these tasks are usually tedious, time-intensive, and cognitively taxing (Tocchetti et al., 2022). One strategy to motivate experts is gamification (Warsinsky et al., 2022), the aim of which is to afford gameful experiences through game design elements (e.g., points, badges) to foster instrumental outcomes like health behavior change or learning effects (Koivisto & Hamari, 2019; Schmidt-Kraepelin et al., 2020).

While the gamification of XAI evaluation seems promising, research on this topic is scarce. Existing studies focus largely on non-expert evaluators (e.g., Leichtmann et al., 2022). As gamification is highly context-sensitive regarding targeted users (Nacke & Deterding, 2017) and experts possess several unique characteristics (e.g., differences in cognitive abilities or low availability (Wienrich et al., 2022)), it is unclear whether insights from these studies are transferable to expert evaluators. Furthermore, the focus of extant studies is often not on gamification but on XAI evaluation: few outcomes of gamification (e.g., motivation) are measured (e.g., Guo et al., 2022), and design rationales that connect presented gamification solutions to relevant problems often lack (e.g., Fulton et al., 2020). Without understanding the relevant problem

URI: https://hdl.handle.net/10125/106542
978-0-9981331-7-1

space, research does currently not offer insights into the inner workings of gamification in expert XAI evaluation tasks and does not support the design of successful gamification concepts in this context (Mulgund et al., 2022). Hence, we seek to explore the problem space of gamified expert XAI evaluation in depth.

To address this goal, we synthesize the relevant literature on gamifying expert XAI evaluation tasks to derive relevant problem space elements. We evaluate these elements through interviews with XAI evaluation experts. In doing so, we make several contributions to extant research. First, by providing insights into the problem space, we lay the bedrock for the development of design knowledge for gamification concepts for expert XAI evaluation tasks. Our results can be used to draw relationships to the solution space (i.e., design solutions to address the identified problems). Second, we provide insights into the peculiarities of experts as a unique user group that is increasingly relevant for gamified systems and contribute to a better understanding of problems to consider when designing gamification concepts for experts. Last, our synthesis of existing gamified XAI evaluation tasks may inspire designers to pick up gamification as a tool to motivate XAI evaluators and support high-quality evaluations.

2. Background

2.1 Explainable AI evaluation

As a response to increasing demands for "unblack boxing" AI models, XAI research aims to produce more transparent and human-understandable AI models while not decreasing model performance (Adadi & Berrada, 2018). XAI approaches can bring benefits like increased trust and transparency in AI models (Lötsch et al., 2021), but also carry several unsolved risks and challenges regarding fairness, accountability, privacy, security and safety, or ethics (Barredo Arrieta et al., 2020). To date, the healthcare domain is by far the most prominent in XAI research, followed by mathematics and biology (Jacovi, 2023). There are two main approaches to XAI: (1) Augment black box AI models to make them interpretable (e.g., by modifying the underlying model architecture) or (2) extend AI models with (post-hoc) explanations to clarify their internal functions (e.g., the relevance of individual features). Most current XAI research focuses on post-hoc explanations (Jacovi, 2023), as these are usually easier to produce (Barredo Arrieta et al., 2020) and possibly adaptable to multiple AI models (Adadi & Berrada, 2018). A current problem is that it still remains unclear what the most appropriate explanation is in a given context for a given individual (Ma et al., 2022). While there exist some functionality-grounded metrics (e.g., runtime, model size, selectivity), there is a need for human-centered evaluation of explanations to ultimately choose the most appropriate ones for the intended users (Zhou et al., 2021). Most current research on human-centered XAI evaluation focuses on non-expert evaluators by applying methods like crowdsourcing (e.g., Jain, 2021). However, because specialized XAI evaluation tasks usually require expertise, effective XAI evaluation often demands involving domain experts with expertise in the relevant XAI application area (Ma et al., 2022). Motivating experts to do XAI evaluation tasks can be challenging due to time constraints, a lack of incentives, and cognitive burden (Tocchetti et al., 2022). In this study, we look at gamification as a way against this drawback.

2.2 Gamification of XAI evaluation tasks

Gamification broadly refers to the use of elements typically found in games (e.g., points, badges, leaderboards) to evoke gameful experiences in non-game contexts (Koivisto & Hamari, 2019). Today, gamification is applied to great effect in domains like education (Koivisto & Hamari, 2019) or healthcare (Schmidt-Kraepelin et al., 2020) to motivate individuals to perform certain behaviors by affording gameful experiences. There is a growing number of studies on experts as a unique user group of gamified systems, investigating, for example, the support of health professionals (Dumitrache et al., 2013) in tasks like data annotation (Warsinsky et al., 2022).

Regarding XAI evaluation tasks, some research exists in the nexus of gamification and AI (Khakpour & Colomo-Palacios, 2020). This research shows that gamification approaches can effectively support AI models but also remarks that these findings are not easily generalizable (Khakpour & Colomo-Palacios, 2020). Overall, this highlights the great potential of gamification for XAI evaluation but also encourages careful design for this context, especially as adding gamification features can jeopardize instrumental evaluation task outcomes (Chu et al., 2020). Some studies on XAI evaluation have started to acknowledge problems related to motivating evaluators (Tocchetti et al., 2022) and accordingly tried to apply game elements to tackle these issues (e.g., Fulton et al., 2020). However, the applied game elements often are not the focus of analysis; very little prescriptive knowledge is provided (e.g., Bansal et al., 2019). Game elements are also often part of games-with-a-purpose (GWAP) environments (Fulton et al., 2020). Compared to a gamified system, a GWAP is usually a full game, which must sacrifice some functionality to uphold its game environment (Liu et al., 2017). Furthermore, these studies mostly look at non-expert evaluators (e.g.,

Leichtmann et al., 2022) and do not consider the unique characteristics of experts. To sum up, the current state of research does not support the informed design of gamification concepts for XAI evaluation tasks.

3. Methods

Broadly, design knowledge may be considered a means-end relationship between problem and solution space. To this end, the problem space often gets neglected in design research (Maedche et al., 2019) despite problem formulation being regarded as the first important step in any research (Van de Ven, 2007). A designer's understanding of a problem can significantly shape how they design artifacts (Maedche et al., 2019). We aim to advance knowledge on the successful design of gamification concepts for expert XAI evaluation tasks by exploring the associated problem space. We conceptualize the problem space via four key concepts, as proposed by Maedche et al. (2019): (1) *stakeholders* (i.e., a person or organization that is involved or affected by the system), (2) *needs* (i.e., the essence of the problem, indicating what is wanted or desired), (3) *goals* (i.e., desired results or a desired state of affairs) and (4) *requirements* (i.e., a condition or capability that must be met or possessed by a system or user). We chose this conceptualization as it provides us with a granular, expressive approach to the problem space (Mulgund et al., 2022). To derive stakeholders, needs, goals, and requirements, we first reviewed extant literature and then evaluated our findings through interviews with experts in XAI evaluation. An overview of our methods can be found in Figure 1.

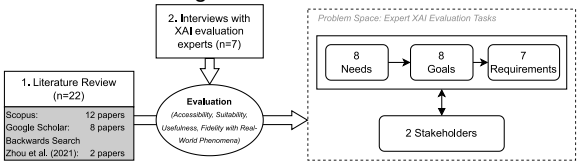

Figure 1. Methods overview

3.1 Literature review

First, we sought to derive initial problem space elements from literature. As the literature on gamifying XAI evaluation tasks is quite scattered, we chose an unstructured approach aimed at breadth. We collected 22 relevant studies by conducting (1) a search in the Scopus database using the search string "*gam* AND explainab* AND AI*" (186 hits last searched on May 29, 2023, 12 relevant papers), (2) a backward search from a recent literature review on XAI evaluation by Zhou et al. (2021) (2 papers), and (3) a Google Scholar search using keywords related to XAI (*interpretable AI, XAI, explainable AI*) and gamification (*gamification, games-with-a-purpose*)(8 papers). While the Scopus search was aimed to cover the major part of peer-reviewed literature, the additional searches served to augment this with possibly cutting-edge research. We made no restrictions regarding publication dates or outlets. Based on titles and abstracts, we included studies where humans are involved in XAI evaluation tasks (i.e., human-centered XAI evaluation, Zhou et al. (2021)) and those tasks were gamified in some way. Detailed information on all relevant papers can be found in SM1 (all material available at: *https://osf.io/357uq*).

We analyzed the 22 relevant papers through thematic analysis following the guidelines by Braun and Clarke (2006). Two researchers independently read the full texts of the papers and assigned initial codes to interesting text passages. They then collated their findings, discussed discrepancies, and iteratively refined the codes, ultimately forming them into themes. For example, we identified many game mechanics like levels, rewards, or points. Making the prior a theme and the latter aspects of that theme. Themes and aspects were ordered in a thematic map. In the end, we included 560 relevant text passages and 169 unique codes grouped into eight themes (cf. section 4.1).

To derive problem space elements, we looked at our current results and tried to identify how themes and subthemes may correspond to stakeholders, needs, goals, and requirements. We organized our results along "problem space chains" following the conceptual model by Maedche et al. (2019). Each chain consists of exactly one need that informs one or more goals that are satisfied by one or more requirements. While forming these chains, we were also constantly on the lookout for relations to gamification, especially how gamification as a solution may help address requirements.

3.2 Expert interviews

The goal of the interviews was to evaluate our identified problem space elements. For this, we needed experts on XAI and XAI evaluation. Hence, potential interviewees were identified via Google and LinkedIn searches and subsequently invited via email. In total, we recruited seven interviewees. All interviewees had at least a Master's degree in computer science or a related field and worked with XAI. Perhaps unsurprising, with a majority of XAI research situated within the healthcare domain (Jacovi, 2023), four of our interviewees had previous experience in the field. All but one had experience with XAI evaluation. Two interviewees had prior work experience with gamification (see also Table 1).

We followed a semi-structured interview approach (for a detailed description, see SM2); this gave us a basic structure required to evaluate our findings and leave our

interviewees with enough room to discuss problem space elements that did not come to our attention while preparing for the interviews (Myers, 2009). After some demographic questions, we asked our interviewees about their experiences with XAI, XAI evaluation, and gamification. We also used this to clear up any conceptual ambiguity and ensure a shared understanding of these concepts.

Table 1. Interviewee characteristics

ID	Occupation	XAI application areas	XAI evaluation-experience*	
			func	**hum**
i01	Developer	multiple	medium	low
i02	Professor	multiple (healthcare, agriculture)	medium	high
i03	Post-doc	multiple	high	high
i04	PhD student	multiple (finance)	low	low
i05	Post-doc & Developer	healthcare	medium	medium
i06	PhD student	multiple (healthcare, climate science)	high	medium
i07	Professor	healthcare	medium	high

*func: functionality-grounded evaluation, hum: human-centered evaluation. As defined by Zhou et al. (2021).

We then successively presented our identified problem space chains to the interviewees and discussed them individually with respect to relevant evaluation criteria (shown in parentheses hereafter). We always started with questions about *accessibility* to ensure a shared understanding. Afterward, we discussed whether interviewees had experienced pertinent needs themselves (*fidelity with real-world phenomena*), whether goals or requirements are suitable to address and satisfy their respective needs or goals (*suitability*), and whether they deem a chain to contain valuable information (*usefulness*). In case of a negative response, we asked for ways to improve the elements in question. We did not discuss every chain with every interviewee; rather, we prioritized discussing chains that fit well with the interviewee's expertise whilst ensuring even coverage of chains across interviews. Using this approach, we noticed that toward later interviews, discussed topics often repeated and no new problem space chains emerged, indicating satisfactory levels of theoretical saturation (Myers, 2009). In the end, we openly asked our interviewees how they would design a gamification concept for expert XAI evaluation tasks. This provided us with valuable information, as interviewees often referenced the discussed problem space chains when providing rationales for their hypothetical designs. Finally, we wrapped up the interview. On average, interviews lasted 57:30 minutes.

To analyze the transcribed interviews, we performed inductive coding to identify text passages that deal with our problem space chains in light of our proposed evaluation criteria (Myers, 2009). At the same time, we also performed open coding on the whole transcripts to identify relevant text passages for our overarching topic. Both codings were initially done by one author and then iteratively discussed and refined within the whole author team. Overall, the accessibility, suitability, and usefulness of our problem space chains were described as high by our interviewees. Each need was also experienced by at least one interviewee in their previous work. Thus, the interviewees mostly confirmed or augmented our findings from the literature. We did, however, make some small adjustments regarding terminology (e.g., ambiguous use of the terms "users" and "experts"). We also added two additional connections between needs, goals, and requirements.

4. Results

4.1 Overview of analyzed literature

Hereafter, we briefly present the eight themes we identified. For a full overview of the thematic map, codes, and frequency matrix, see SM3.

Evaluators. The literature identifies two major evaluator groups: non-experts (Ma et al., 2022), which are usually available and often mentioned as related to crowdsourcing, and domain experts (Zhou et al., 2019), which are less available but important for use cases where expertise is essential (Butz et al., 2022).

Evaluation task. We identified several tasks evaluators might perform to gain insight into the human understandability of an explanation, including comparing or ranking multiple explanations (Butz et al., 2022; Sevastjanova et al., 2021), rating explanations (Newn et al., 2019; Schlippe & Sawatzki, 2022), deleting unhelpful (Sevastjanova et al., 2021), or creating new explanations (Suryanarayana et al., 2022).

Explanation complexity. The complexity of explanations is often cited as a source of cognitive load for evaluators (Jain, 2021). In the context of XAI evaluation, studies discuss problems related to task difficulty (Lu et al., 2021), presentation of explanations, and effective task selection (Ben David et al., 2021).

Explanation quality. Regarding explanation quality, literature frequently emphasizes a lack of ground truth and resulting uncertainty for experts and developers (Fulton et al., 2020). Studies discuss several explanation quality criteria (e.g., usefulness (Ray et al., 2019)) and quality control (e.g., consensus mechanisms (Leichtmann et al., 2022)).

Dual outcomes of gamification. The literature discusses various instrumental outcomes when utilizing gamification for XAI evaluation, like developing a mental model of AI (Bansal et al., 2019) and a need for creativity (Fulton et al., 2020). Yet, most focus on experiential outcomes, like how to motivate users

(Shingjergji et al., 2022) and hold user engagement (Berger & Müller, 2021) through gamification.

Game mechanics. Various gamification elements were embedded into XAI evaluation tasks in our reviewed literature: scores/points (Chu et al., 2020), rewards, badges, challenges, levels/progression (Sevastjanova et al., 2021), leaderboards (Tocchetti et al., 2022), and narratives (Ehsan et al., 2018).

Incentives. A large part of the literature revolves around how to provide incentives to evaluators. While monetary incentives exist (e.g., crowdsourcing), most research focuses on non-monetary social mechanisms (e.g., teamwork or competition (Nazir et al., 2023)).

Feedback loop. Literature often discusses bilateral feedback between the XAI developer and evaluator. Feedback from the evaluator to the developer allows them to express their domain knowledge (Guo et al., 2022), while feedback from the XAI developer to the evaluator can help them gauge their performance or provide guidance for unclear evaluation tasks (Sevastjanova et al., 2021).

Some other discourses that did not fit or merit an entire theme include designing GWAPs centered around XAI evaluation (Fulton et al., 2020), explainability human-AI collaboration (Ehsan et al., 2018), or how to avoid automation bias (Wienrich et al., 2022).

4.2 Problem space of gamified expert XAI evaluation

Based on our combined insights from the literature and the interviews, we identified two stakeholders and several problem space chains consisting of eight needs, eight goals, and seven requirements (see Figure 2). We identified two main stakeholders: First, an XAI developer is any person or organization that aims to develop an (explainable) AI solution for a given use case, for which they may then require an evaluation of the included explanation. Thus, we consider XAI developers to be the XAI evaluation task provider. The second stakeholders are domain experts, which are individuals who are proficient with the tasks the XAI system is intended to solve for a given use case; that is, they are knowledgeable about the data used, possible output predictions, and capable of solving the task manually (Sevastjanova et al., 2021). The domain experts constitute the prospective users of the finished XAI solution (Ma et al., 2022) and the main group that the gamification concepts seek to motivate.

While these two groups emerged as the main stakeholders for expert XAI evaluation, there may be further affected parties, for example, customers or patients relying on the decision supported by the relevant XAI solution (Bansal et al., 2019). However, as these parties are not directly involved in the evaluation and are not the target of the gamification concepts, we largely omit them from future discussion.

Need 1: Experts are unmotivated to participate in XAI evaluation. Domain experts are often highly occupied with their work, especially in high-stakes contexts like healthcare or finance (Butz et al., 2022; Suryanarayana et al., 2022). Thus, XAI evaluation tasks offer little intrinsic motivation for experts: They are usually tedious and time-intensive (Sevastjanova et al., 2021; Tocchetti et al., 2022) and offer little inherent benefit if they are separate from the evaluator's core work. A goal we identified is to ***motivate experts to partake in evaluation tasks*** **(G1)**. The associated requirement is to ***provide incentives*** **(R1)**. Monetary incentives work well for non-expert evaluators (Ben David et al., 2021) but less so for experts due to their relatively high cost (Suryanarayana et al., 2022). As one interviewee said: "*So you can't pay them for participating [...]. You have to rely on their willingness to participate or an obligation from their boss or something"* (i03). Hence, non-monetary incentives may be applied, like communicating the value of a XAI evaluation task to evaluators: "*if [evaluators] see the value and if they are involved in the project, I think there's a high chance that they participate in evaluations*" (i03). Similarly, a system can highlight how participation can educate experts and help improve their own competence (Tocchetti et al., 2022). One interviewee was skeptical about gamification in these roles: "*I think there is very little gamification-driven incentive that could be there to motivate them. To join it, [explain to the expert]: what are you gaining out of this?*" (i05). Literature, however, cites social mechanisms like teamwork or competition as promising incentives (Chu et al., 2020) that are often core parts of gameful designs. Another relevant incentive is appealing to humans' desire for entertainment and fun (Chu et al., 2020) by using gamification to turn a tedious task into something fun. One interviewee remarked: "*[Gamification] is making [the evaluation task] easy to access [...] And making it fun is also something it should be*" (i06). The same interviewee goes on to caution that too many gamification elements can also tilt motivation by clashing with the seriousness associated with XAI evaluation: "*I would say it shouldn't be too gamified because it's still domain experts, right? They take themselves seriously.*" (i06). Overall, the role of gamification as an incentive remained contested.

Need 2: Evaluation tasks are not engaging for experts. Literature concurs that XAI evaluation tasks are tedious and repetitive, which necessitates a mechanism to keep evaluators engaged once they have started to work on the task (Fulton et al., 2020) to avoid them abandoning a task altogether. Dropouts due to low engagement are particularly dire with respect to experts

being already hard to access for many XAI developers (Suryanarayana et al., 2022). Therefore, ***minimizing dropouts*** **(G2)** is an important goal. The previously mentioned ***incentives*** **(R1)** contribute to this goal when providing incentives for successfully finishing a task (Ma et al., 2022). One interviewee remarked how they preferred gamification over other incentives: *"Incentives are not working in certain aspects. It is better to have this flow, to have this engagement. I also think a gamification approach would be the most efficient one for keeping them in the flow"* (i02). Another requirement we found regarding this goal is to ***select expert tasks in an effective and efficient way*** **(R2)**. This includes splitting evaluation tasks into granular parts (Suryanarayana et al., 2022) to avoid overwhelming evaluators by unlocking content as needed. Levels can be a useful gamification element here, as they "provide small goals that engage experts to keep striving to reach the next one" (Sevastjanova et al., 2021, p. 6). Effective task selection also includes minimizing bothersome workload and shortening the evaluation task as much as possible (Guo et al., 2022). Gamification elements may lengthen system interaction times, which is important to consider when experts are only available for a certain time: "*It is just time […] every expert has a lot of work to do. [...] if they have to wait, they will reduce their engagement, they will reduce their motivation, and then they likely dropout*" (i02).

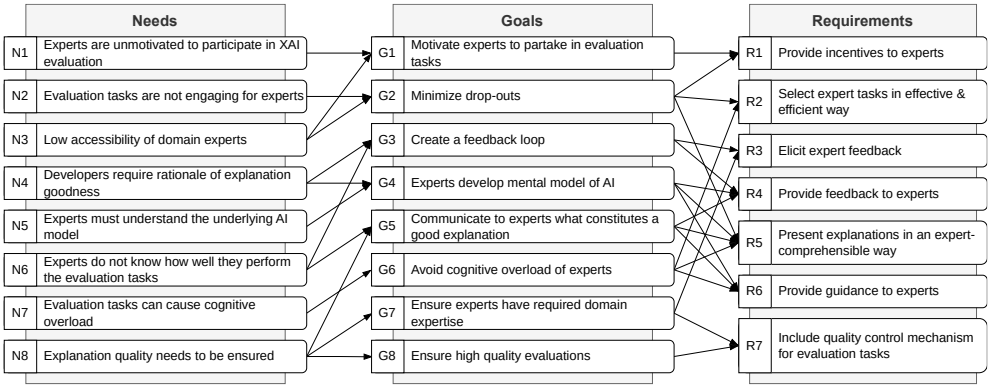

Figure 2. Overview of identified needs, goals, and requirements

Need 3: Low accessibility of domain experts. Our reviewed literature widely recognizes insufficient availability of domain experts to XAI developers (Butz et al., 2022). Therefore, many resort to non-expert evaluators via approaches like crowdsourcing (Shingjergji et al., 2022). Some studies specifically include game environments to try and alleviate the need for expertise (Guo et al., 2022) or to improve scalability of the evaluation (Chu et al., 2020). However, studies recognize problems with non-expert evaluators, such as a gap of understanding (Shingjergji et al., 2022). Thus, in XAI evaluation, experts are still very difficult to replace, especially in high-stakes decision-making contexts (Lu et al., 2021). This need accentuates our previous deliberations on motivating and engaging XAI evaluators. ***Motivating evaluators*** **(G1)** gets even more crucial when the population of possible domain experts is small; so does keeping evaluators engaged (cf. Need 2) and ***minimizing drop-outs*** **(G2)** to avoid losing precious domain experts: *"This is one of the major goals to minimize dropouts. The worst case is if you [lose participants]"* (i02).

Need 4: Developers require rationale of explanation goodness. The "goodness" or human understandability of an explanation may be assessed by rating, comparing, ranking explanations (Wienrich et al., 2022), or using an explanation to modify the underlying AI model, also referred to as "break the bot" (Lu et al., 2021). To this end, providing experts with performance scores (points) lends itself well to identifying the value of different explanations in a playful manner (Jain, 2021). However, XAI evaluation tasks often lack an underlying ground truth and are thus a subjective and creative process that may even involve intuition (Newn et al., 2019). Hence, to gain a better understanding of explanations, XAI developers do not only need to know how an explanation was evaluated but also why it was evaluated in a certain way. The corresponding goal is the ***creation of a feedback loop*** **(G3)** that is addressed by both ***providing feedback to experts*** **(R4**, cf. Need 5) and, more importantly for this need, ***eliciting feedback from the experts*** **(R3)**. Feedback can, for example, be elicited by having evaluators provide textual justifications for their explanation rankings (Ehsan et al., 2018). To this end, social gamification elements can allow experts easier communication with XAI developers: *"I think social interaction [...] with the developer [...] you can directly write them. It's so much more familiar, […] and make people feel less overwhelmed"* (i06). The elicited feedback may serve as an interface into domain experts' knowledge (Sevastjanova et al., 2021). One important aspect to consider is that the impact of providing feedback may

not be transparent to the expert (Guo et al., 2022). Here, gamification can be beneficial to encourage experts to provide rationales for their evaluations, for example, by giving out points for doing so (Jain, 2021), or by increasing the perceived value of feedback through a narrative (Butz et al., 2022).

Need 5: Experts must understand the underlying AI model. XAI evaluation is a creative process where evaluators require some understanding of the capabilities of the underlying AI model (Bansal et al., 2019). Therefore, a goal is that ***experts develop a mental model of the underlying AI*** **(G4)**. Such a mental model can, for example, include how the AI will predict for a given input. Mental models can, however, be difficult to develop for evaluators, as they have only a limited number of interactions with an AI (Bansal et al., 2019). Therefore, studies emphasize the importance of ***presenting explanations in an expert-comprehensible way*** **(R5)**. Visualizations and gamification of explanations may help here (Guo et al., 2022). However, benefits may be lost if evaluators are not able to utilize the information given to them due to inexperience (Newn et al., 2019). It is important to consider that experts may have different levels of expertise (Tocchetti et al., 2022): "*[A chess] Grandmaster would understand immediately. [...] but to a test player, you need to provide a different kind of explanation. So that's very [user-dependent]*" (i01).

Another requirement for mental model development is to ***provide guidance to experts*** **(R6)**, that is, making suggestions on appropriate views or next steps. Guidance is particularly important at the start of a system interaction in order to avoid overwhelming experts. Gamification mechanics like tutorials or content unlocking can be useful here to introduce system functions in a successive manner. During system interaction, pre-defined badges may also be useful to guide experts toward possible tasks (Sevastjanova et al., 2021). One interviewee also suggested the use of avatars as a guidance persona like Microsoft's office assistant "Clippy": *"The paper clip. [The avatar] guiding you through the steps of what's going on and so you can interact with that"* (i06). Lastly, ***providing feedback to the experts*** **(R4)** as part of the proposed ***feedback loop*** (cf. **G3**) also contributes towards mental models, as evaluators can gradually improve their mental model based on the feedback they are given (Bansal et al., 2019).

Need 6: Experts do not know how well they perform the evaluation tasks. Expert evaluators usually want to gauge their own performance and see the results of their actions in order to feel in control (Guo et al., 2022). However, as XAI evaluations often rely on subjective opinions of experts due to the lack of ground truth (Fulton et al., 2020), it is often difficult to measure performance objectively (Jain, 2021). This problem founds the second part of our proposed ***feedback loop*** **(G3)**, particularly the part of ***providing feedback to experts*** **(R4)**. Game environments can be specifically useful, as they help experts to operate under uncertainty. Badges or achievements may serve as representations of user success, while scores can be used as a metric of performance that may also be compared across evaluators (Sevastjanova et al., 2021): *"You can always have rewards, and incentives. Challenges, quests, for example. Progress tracking. I would say this all is a kind of feedback"* (i02).

While expert evaluators have deep knowledge of their domain, they are not necessarily aware of what XAI developers in said domain require (Tocchetti et al., 2022): *"Experts often do not know exactly what the goal of the evaluation is […] and they do not really know what you want to know. […] therefore, this feedback is very important."* (i02). Hence, to set a quality standard, an important goal is to ***communicate to experts what constitutes a good explanation*** **(G5)**. Literature entails quality criteria for explanations like intuitiveness, robustness, or perceived helpfulness (Ray et al., 2019). The importance of these may be communicated when ***providing guidance to experts*** **(R6)** and ***providing feedback to experts*** **(R4)**. From a gamification perspective, tutorials can provide a low-pressure environment to allow experts to gain confidence in understanding explanation quality (Newn et al., 2019). Providing high-quality evaluations can also be rewarded with status-based awards to signify their goodness (Tocchetti et al., 2022). Lastly, ***presenting explanations in an expert-comprehensible way*** **(R5)** communicates what a good explanation may constitute.

Need 7: Evaluation tasks can cause cognitive overload. XAI evaluation tasks often require evaluators to process a lot of information simultaneously (e.g., AI model, output prediction, and explanation), which can easily lead to cognitive overload that can, in turn, decrease task engagement (Sevastjanova et al., 2021). Hence, ***avoiding cognitive overload of experts*** **(G6)** is an important goal. The engagement of evaluators can drop if they are faced with a data overload in an evaluation task. ***Providing guidance to experts (R6)*** can feather this engagement drop, in particular by unlocking content or information in a controlled and successive manner (Lu et al., 2021) to allow users to easily contextualize newly revealed information. Gamification can support this by unlocking content as a reward (Sevastjanova et al., 2021).

Another important factor in avoiding cognitive overload is to omit irrelevant information (i.e., decrease task dimensionality (Bansal et al., 2019)). To this end, ***selecting tasks in an effective and efficient***

***manner* (R2)** helps to ensure that experts are always on a suitable task for them: *"Minimize the information, minimize the number of questions asked, or minimize the number of tasks you have."* (i03). Additionally, ***presenting explanations in an expert-comprehensible way* (R5)** includes choosing a suitable explanation gestalt. With respect to processing of information, research acknowledges that gamification elements can ease processing of information but also highlights that gamification elements can pose as additional distracting stimuli (Newn et al., 2019). Thus, gamification elements may be discordant with the idea of keeping things simple: "*Yeah, it should not be, like, stressful. I mean, of course, it's going to be to some level, but it shouldn't exponentially increase [...] So I would say, make it as simple as possible*" (i05).

Need 8: Explanation quality must be ensured. XAI research ultimately seeks to produce useful explanations to improve the understandability of AI models (Barredo Arrieta et al., 2020). Hence, the intuitively associated goal is to ***ensure high-quality evaluations* (G8)**. We did identify one new requirement to address this goal, which is to ***include quality control mechanisms for evaluation tasks* (R7)**. Literature features several mechanisms, most notably consensus mechanisms (Butz et al., 2022) and peer-review systems (Tocchetti et al., 2022). Applying gamification, scores can be used as a metric for evaluation quality (Shingjergji et al., 2022), while social mechanisms can motivate evaluators to improve existing evaluations from other evaluators. Gamification may also serve as an inherent quality mechanism when "providing accurate [evaluations] is the ideal way to play this game" (Fulton et al., 2020, p. 4). Further, interviewees remarked that experts are "*[…] very competitive, and they want to be the best at their job*" (i06). Therefore, allowing experts to compete over the best evaluation can be a quality control mechanism (Ray et al., 2019).

Regarding evaluation quality, the reviewed literature highlights that higher evaluator expertise can lead to higher levels of explanation understanding (Wienrich et al., 2022). Therefore, another goal is to ***ensure that expert evaluators have the required domain expertise* (G7)**. This goal is addressable by assessing an evaluator's expertise at the start of an evaluation task, which we view as a form of ***eliciting feedback from experts* (R3)**. Grouping evaluators into different levels and allowing them to rise through levels may even act as motivation for evaluators (Tocchetti et al., 2022). ***Quality control mechanisms* (R7)** may also contribute to identifying evaluators with insufficient expertise. Shingjergji et al. (2022), for example, propose a gamification design where low-quality evaluations are resembled by low scores.

5. Discussion

5.1 Principal findings

Our results reveal that expert XAI evaluation is a unique context for gamification with several promising roles that gamification may fulfill. First, the literature and interviewees were generally optimistic about the benefits of gamifying expert XAI evaluation tasks. We found gamification to fit better into some problem space chains than others. For example, scores fit very intuitively as incentives (cf. R1), performance criteria (cf. N6, R7), or as feedback mechanisms (cf. G3, R4, R6); so did levels to lower cognitive load (cf. G6) or to group evaluators by expertise (cf. G7). It was, however, difficult to find anchor points for gamification in chains that were intrinsic to XAI evaluation, like low accessibility of experts (cf. N3) or presentation of explanations in a comprehensible way (cf. R5). Our interviewees often remarked that to be effective, gamification for expert XAI evaluators requires a separate foundation to build on, like payment or inherent benefits from the evaluation for evaluators. Thus, we view gamification as an augmentor for expert XAI evaluation.

Second, it was interesting to see that only some problems associated with expert XAI evaluation tasks were rooted in the XAI evaluation task itself (e.g., the high cognitive load), while most were rooted in unique characteristics of experts as a user group (e.g., their low accessibility). This is also interesting for current gamification research, which often emphasizes the context-sensitivity of gamification (Nacke & Deterding, 2017). The existence of problem space chains only rooted in characteristics of experts suggests that, in this case, experts as a user group become the key contextual factor. This could mean that insights between different gamification application areas are transferable as long as the factor "experts as users" stays consistent. Compared to studies on the design of gamification concepts to involve experts in other tasks (Dumitrache et al., 2013), our findings also reveal how AI as a contextual factor may influence the needs of experts (e.g., having to understand complex AI models). Therefore, our findings provide unique insights into how gamification can shape expert-AI collaboration.

Lastly, research on XAI evaluation seems quite fragmented across different domains. While highlighting the importance of conducting XAI evaluations, it stood out to us that there exist few standards on how to properly evaluate explanations, with or without experts (Zhou et al., 2021). In terms of gamification, we observed that research mostly opted for structural gamification elements (i.e., elements

added to the structure of the content without altering the content (Koivisto & Hamari, 2019)). Only rarely did we identify more complex elements like narratives (e.g., Butz et al., 2022). On one hand, researchers may worry that gamifying their system could jeopardize the evaluation, such as when gamification elements distract from the evaluation task (Newn et al., 2019). On the other hand, designers may be reluctant to accept the comparatively high design efforts of non-structural gamification elements, as research on their benefits for XAI evaluation is limited. Finally, most studies striving to gamify XAI evaluation focus on simple, non-expert scenarios and avoid high-stakes decision contexts like healthcare, though exceptions exist (e.g., Butz et al., 2022; Nazir et al., 2023)

5.2 Implications

For research, our findings contribute to a better understanding of problems associated with expert XAI evaluation and how gamification can contribute to solving them. Most importantly, our results indicate that gamification is a promising solution for many of the problems associated with expert XAI evaluation but may not be able to comprehensively solve them all. Our findings also highlight that even though it may clash with the inherent seriousness that experts bring with them, gamification can indeed be valuable to motivate and engage experts. However, experts should be considered a unique user group for gamified systems that, accordingly, require a unique gameful experience (Warsinsky et al., 2022). Our approach also shows the value of proper problem formulation. By breaking down expert XAI evaluation into problem space chains consisting of stakeholders, needs, goals, and requirements, we were able to achieve a granular understanding and, hence, comprehensively describe the problem space. To this end, our findings may inform the development of (gamification) solutions for these problems. By connecting prospective gamification concepts to the individual problem space chains, researchers can achieve a better understanding of the inner workings of said concepts.

For practice, our findings highlight that gamification concepts for expert XAI evaluation tasks require careful design based on the unique characteristics of XAI evaluation and experts. Designers should carefully choose which problem space chains to tackle via gamification and which to repudiate. They should consider that gamification may not be a comprehensive solution but rather an augmentor that requires a separate foundation to be effective. Overall, our results can give designers of gamification concepts for expert XAI evaluation tasks a better idea of which problems are important to tackle.

5.3 Limitations and future research

A limitation of our study is that we only focused on the problem space. Therefore, we did not produce complete design knowledge in the sense of means-end relationships between problem and solution space (Maedche et al., 2019). In their current state, our results are not immediately actionable for designers of gamification concepts for expert XAI evaluation tasks. A logical next step would be the implementation and evaluation of a gamified system. Doing so does not only help to understand the solution space but can also lead to a better understanding of the underlying problems. We also acknowledge the exploratory nature of our study. We did not conduct a systematic, comprehensive literature search, and we had comparatively few interviewees (n=7) who were also only XAI developers. While we think that our triangulation of extant literature and expert interviews led to a high degree of theoretical saturation, some topics emerged that we were unable to investigate further due to a lack of data (e.g., other stakeholder experts or concerns about automation bias). Future research may be able to tease out more about these topics by applying more systematic, comprehensive approaches (e.g., a systematic literature search).

6. Conclusion

In this study, we conducted a problem space exploration on gamification concepts for expert XAI evaluation tasks. Based on 22 extant studies and seven expert interviews, we derived two stakeholders, eight needs, eight goals, and seven requirements. Our results suggest that gamification can improve motivation and engagement among expert XAI evaluators, but also indicate concerns regarding the effectiveness of gamification in this domain. We hope our findings can support design and implementation of gamification concepts for expert XAI evaluation in practice.

7. References

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138-52160.

Bansal, G., Nushi, B., ..., & Horvitz, E. (2019). *Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance*. HCOMP 2019, Stevenson, USA.

Barredo Arrieta, A., Díaz-Rodríguez, N., ..., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inform Fusion*, *58*, 82-115.

Ben David, D., Resheff, Y. S., & Tron, T. (2021). *Explainable AI and Adoption of Financial Algorithmic Advisors: An Experimental Study*. AIES ‘21, Virtual.

Berger, F., & Müller, W. (2021). Back to Basics: Explainable AI for Adaptive Serious Games. In *Serious Games* (67-81).

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qual Res Psychol*, *3*(2), 77-101.

Butz, R., Schulz, R., Hommersom, A., & van Eekelen, M. (2022). Investigating the understandability of XAI methods for enhanced user experience: When Bayesian network users became detectives. *Artif Intell Med*, *134*, 102438.

Chu, E., Gillani, N., & Priscilla Makini, S. (2020). *Games for Fairness and Interpretability*. Web Conference, Taipei, Taiwan.

Dumitrache, A., Aroyo, L.,..., & Levas, A. (2013). Dr. Detective: combining gamification techniques and crowdsourcing to create a gold standard for the medical domain. In *Crowdsourcing the Semantic Web*.

Ehsan, U., Harrison, B., Chan, L., & Riedl, M. O. (2018). *Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations*. AIES '21, New Orleans, LA, USA.

Fulton, L., Lee, J.,..., & Perer, A. (2020). *Getting Playful with Explainable AI: Games with a Purpose to Improve Human Understanding of AI*. CHI 2020, Oahu, USA.

Guo, L., Daly, E. M.,..., & Knijnenburg, B. (2022). *Building Trust in Interactive Machine Learning via User Contributed Interpretable Rules*. IUI '22, Helsinki, FI.

Jacovi, A. (2023). *Trends in Explainable AI (XAI) Literature*. Retrieved June 8, 2023 from arxiv.org/abs/2301.05433v1.

Jain, M. (2021). *Crowd-Sourced Evaluation of Explainable AI Techniques with Games* [CMU Pittsburgh, USA].

Khakpour, A., & Colomo-Palacios, R. (2020). Convergence of Gamification and Machine Learning: A Systematic Literature Review. *Technology, Knowledge and Learning*, *26*(3), 597-636.

Koivisto, J., & Hamari, J. (2019). The rise of motivational information systems: A review of gamification research. *Int J Inform Manag*, *45*, 191-210.

Leichtmann, B., Hinterreiter, A.,..., & Mara, M. (2022, September 21). *Explainable Artificial Intelligence improves human decision-making: Results from a mushroom picking experiment at a public art festival.*

Liu, D., Santhanam, R., & Webster, J. (2017). Toward Meaningful Engagement: A Framework for Design and Research of Gamified Information Systems. *MIS Quarterly*, *41*(4), 1011-1034.

Lötsch, J., Kringel, D., & Ultsch, A. (2021). Explainable Artificial Intelligence (XAI) in Biomedicine: Making AI Decisions Trustworthy for Physicians and Patients. *BioMedInformatics*, *2*(1), 1-17.

Lu, X., Tolmachev, A., ..., & Kashima, H. (2021). Crowdsourcing Evaluation of Saliency-Based XAI Methods. In *Machine Learning and Knowledge Discovery in Databases* (431-446).

Ma, H., McAreavey, K., McConville, R., & Liu, W. (2022). *Explainable AI for Non-Experts: Energy Tariff Forecasting*. ICAC, Birmingham, GB.

Maedche, A., Gregor, S., Morana, S., & Feine, J. (2019). *Conceptualization of the Problem Space in Design Science Research*. DESRIST 2019, Worcester, USA.

Mulgund, P., Purao, S., & Agrawal, L. (2022). *Fathers with Postpartum Depression: A Problem Space Exploration*. DESRIST 2022, St. Petersburg, USA.

Myers, M. D. (2009). *Qualitative research in business & management*. Sage Publications Ltd.

Nacke, L. E., & Deterding, S. (2017). The maturing of gamification research. *Comp Hum Behav*, *71*, 450-454.

Nazir, S., Dickson, D. M., & Akram, M. U. (2023). Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. *Comput Biol Med*, *156*, 106668.

Newn, J., Singh, R.,..., & Vetere, F. (2019). Designing Interactions with Intention-Aware Gaze-Enabled Artificial Agents. In *Human-Computer Interaction – INTERACT 2019* (255-281).

Ray, A., Yao, Y.,..., & Burachas, G. (2019). *Can You Explain That? Lucid Explanations Help Human-AI Collaborative Image Retrieval*. HCOMP 2019, Stevenson, WA, USA.

Schlippe, T., & Sawatzki, J. (2022). AI-Based Multilingual Interactive Exam Preparation. In *Innovations in Learning and Technology for the Workplace and Higher Education* (396-408).

Schmidt-Kraepelin, M., Toussaint, P. A.,..., & Sunyaev, A. (2020). Archetypes of Gamification: Analysis of mHealth Apps. *JMIR Mhealth Uhealth*, *8*(10), e19280.

Sevastjanova, R., Jentner, W.,..., & El-assady, M. (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.

Shingjergji, K., Iren, D.,..., & Klemke, R. (2022). *Interpretable Explainability in Facial Emotion Recognition and Gamification for Data Collection*. ACII 2022, Nara, Japan.

Suryanarayana, S. A., Sarne, D., & Kraus, S. (2022). Explainability in Mechanism Design: Recent Advances and the Road Ahead. In *Multi-Agent Syst* (364-382).

Tocchetti, A., Corti, L., Brambilla, M., & Celino, I. (2022). EXP-Crowd: A Gamified Crowdsourcing Framework for Explainability. *Front Artif Intell*, *5*, 826499.

Van de Ven, A. H. (2007). *Engaged Scholarship: A Guide for Organizational and Social Research.* Oxford University Press.

Warsinsky, S., Schmidt-Kraepelin, M.,..., & Sunyaev, A. (2022, June 1–3). *Gamified Expert Annotation Systems: Meta-Requirements and Tentative Design*. DESRIST 2022, St. Petersburg, FL, USA.

Wienrich, C., Carolus, A., Roth-Isigkeit, D., & Hotho, A. (2022). Inhibitors and Enablers to Explainable AI Success: A Systematic Examination of Explanation Complexity and Individual Characteristics. *Multimodal Technologies and Interaction*, *6*(12).

Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics*, *10*(5).

Zhou, J., Hu, H.,..., & Chen, F. (2019). Physiological Indicators for User Trust in Machine Learning with Influence Enhanced Fact-Checking. In *Machine Learning and Knowledge Extraction* (94-113).