CardLab: A Simple Co-Creative Interface for Designing and Testing Cards in Hearthstone

Alexander Elton-Pym and **Kazjon Grace**
Designing with AI Lab
The University of Sydney
[alex.elton-pym, kazjon.grace]@sydney.edu.au

Abstract

Designing content for multiplayer competitive strategy games, such as collectible card games, is a complex process. Creating original, fun, and balanced content can be particularly challenging in real-world game contexts. This paper presents a pilot study of CardLab, a user-friendly creative interface for card creation and testing in (a constrained version of) the digital card game *Hearthstone*. Our study explores how designers responded to the system's feedback based on simulated games. CardLab aims to help designers more rapidly create high-quality cards, while reducing the need for extensive playtesting.

Introduction

Designing new content is a regular and essential component of many competitive strategy games, especially collectible card games (CCGs). Players expect new content to be regularly released to keep the game fresh, offer new strategic challenges, and shake up existing metagames (Carter, Gibbs, and Harrop 2012). However, keeping this new content original, fun, and balanced is a time-consuming and challenging task, often requiring designers to spend countless hours playtesting new content. As more content is released over the life of a game, generating original ideas becomes increasingly difficult. For example, in Blizzard's popular CCG *Hearthstone* (2014), over 300 new cards are released each year, requiring designers to work quickly to design, develop, and test each card before public release. Balancing content can be particularly challenging as the ecosystem of existing strategies grows, forming exponentially more interactions, any of which could lead to an unintentionally powerful combination. While balance is only one component of what makes a game fun, it is critical in competitive strategy games (Hoover et al. 2020).

As an approach to simplifying this huge and ongoing design problem (specifically in *Hearthstone*, but in concept for any competitive strategy game), we propose using a simulation engine to automatically playtest new cards, coupled with a simple user interface that allows designers to rapidly prototype them. Currently, designers and players regularly rely on statistical inference from game replay data to determine the performance of cards. This data exists only for existing cards, not new and proposed ones. Our system provides accurate simulation data accompanied by a variety of additional behavioural variables, such as the change in average game length. Our simple visual interface enables designers to prototype cards quickly without any need for coding, and the simulation engine allows for detailed feedback on their efficacy and characteristics. Card simulation is a computationally expensive task, and the feedback is only available after several hours — but this is still orders of magnitude less than it would take to playtest with humans. In one sense, our approach is intended as a speculation on the future of simulation-assisted co-creativity: to what degree can large-scale game simulation help game designers with creative tasks? We propose that this more-rapid feedback loop between design, analysis, and iterative re-design, forms a new kind of co-creative system for game content design that will become only more effective as the cost of compute decreases.

We conducted a pilot study with several *Hearthstone* players, each of whom designed a set of nine cards, which were in turn tested through 3,000 simulated games on the popular *Hearthstone* simulator *Spellsource*[1]. In a follow-up session, each user received feedback on the performance and behaviour of their cards and had the opportunity to make modifications to their designs. We analysed the participants' responses to feedback on each card's behaviour, as well as how they accordingly modified their designs.

Background

Competitive strategy games are an age-old form of entertainment, competition, and research. Games like chess have been used to benchmark human and computer capabilities since their inception. The vast majority of these games change very slowly, if at all over the years. For example, our modern version of chess has mostly remained unchanged since its 10th-century origin. This contrasts starkly with modern video games, where increasingly, game studios are following a live service model where games are continuously updated with new content throughout their lifecycle (Dubois and Weststar 2022). In this live service model, retaining customers is key to the economic success of the game, and providing a steady stream of new content while maintaining game balance is often seen as one of the most

[1]Berman and Gale, github.com/hiddenswitch/Spellsource

important factors for player's experience (Adams 2014).

Computationally creative systems have been shown to be able to assist game designers in creating balanced, fun, and original content for their games. Tanagra is a key early example of mixed-initiative creative game content design, constructing platformer levels with a focus on validating playability and comparing different generators for their expressivity (Smith, Whitehead, and Mateas 2010). Sentient Sketchbook (Liapis, Yannakakis, and Togelius 2013) is another creative interface that enables designers to collaborate with an AI to design levels for real-time strategy games. Baba is Y'all (Charity, Khalifa, and Togelius 2020) is notable in our context for incorporating automated playtesting.

Blizzard's collectible card game *Hearthstone* (2014) is – like many large, popular, continually updated games – a highly complex design domain. Evidence for this can be found in the many volatile online discussions about the relative power of new content. Research on Hearthstone has included developing adversarial agents (Świechowski, Tajmajer, and Janusz 2018) and balancing existing metagames (de Mesentier Silva et al. 2019). Like its spiritual grandparent game *Magic: The Gathering*, the space of possible *Hearthstone* decks is astronomically large, making it computationally prohibitive. The scale of this challenge has led to many simulation, analysis, and archive-exploration tools being developed by both the player community and academic researchers (Dockhorn and Mostaghim 2019). Of particular note are experiments in using neural surrogate models to predict game outcomes in reduced time (compared to extensive simulation) (Zhang et al. 2022), but it remains to be seen whether such surrogate approaches could be extended to work with new cards.

CardLab prototype

Our prototype creative interface, CardLab, enables designers to rapidly create simple *Hearthstone* cards (Figure 1). To simplify the space of required simulations (but still keep with the complexity and spirit of *Hearthstone*), we posit a "miniature" version of the game, consisting of only basic cards from the classic set. We further restrict this version to include only Hunter, Warrior, and Mage classes and require all decks to consist of 15 pairs of cards; we designate this format "classic lowlander". In CardLab, users are able to design minion cards with keywords, custom stats (mana cost, attack, and health), and simple "battlecry" effects. This simple prototype allows us to test cards in a "mini-metagame" of the three basic class decks from the game's practice mode.

After cards are converted into a format readable by *Spellsource* and inserted into the basic decks, they are simulated in 1000 games each versus Hunter, Warrior, and Mage. Simulating fewer games was observed to potentially obscure the effects of subtle card changes, while simulating more did not tend to reveal more effects. With this many games, each card takes around 40 minutes to simulate on a single 2022-era high-end workstation, making it infeasible for the interface to provide feedback online, necessitating a follow-up session. Our experiments with high-performance computing indicate that it's feasible to provide live-updating feedback after only a minor delay. We select the three most statistically significant behavioural statistics (using a T-test comparing against baseline decks) along with winrates (using Bernoulli trials) and present these to the user.

Figure 1: The CardLab interface, enabling designers to make comparisons to existing cards. A version of CardLab is hosted at `hearth-mici.web.app`

Study protocol

The user study consists of two half-hour sessions. In the first session, users design three cards for Hunter, Warrior, and Mage, while describing their design choices and thought process. Users' level of expertise with card games is determined by asking them to describe their history with *Hearthstone* and other CCGs. Throughout the study, users are prompted with questions such as "What role do you see this card filling in a deck?" to assist them in thinking aloud.

In the second session, conducted 1-7 days later, users received statistics on each card's performance in simulated games. The simulated decks are direct copies of *Hearthstone's* basic decks, with 2 copies of a vanilla neutral (i.e. non-class-specific, low-cost, no special abilities) minion selected for substitution. Cards are simulated with the *Spellsource Hearthstone* simulator, a popular java-based simulator. Games are played using a default AI from *Spellsource* which uses a form of the Minimax algorithm. This heuristic scores the hypothetical game state that would result from taking each possible move, with a policy that has been optimised with an evolutionary approach.

We asked each user to comment on their cards' performance and behaviour after receiving the feedback from the simulator, and if the results were expected or surprising. We also asked if and how they wanted to modify their cards, categorising their choices as no modification, minor modification, or significant modification. We conducted a thematic analysis of the think-aloud and post-session interviews in order to explore the design motivations of our users and the way they were affected by the simulation results. A large language model (GPT 3.5) was used (in parallel with human coding) as a supportive aid in the first pass of coding, but the final decision for all categories was human.

Figure 2: An example of the reported simulation results, included are the winrates and the most significant behavioural statistic for each match-up.

Results

We conducted our user study with 8 total users. These users have a variety of levels of experiences with *Hearthstone* and other CCGs. Two users were novices who played through the *Hearthstone* tutorial and for an hour with the basic Hunter, Warrior, and Mage decks. Four users were intermediate players who had moderate experience, with most playing when *Hearthstone* was first released. Two users had extensive experience with *Hearthstone*, having played during multiple expansions as well as experience with other CCGs like *Magic: The Gathering* and *Legends of Runeterra*. Here we present both a thematic analysis of their design motivations during the card design/re-design task, as well as an analysis of how they responded to our system's simulation results predicting their cards' performance and behaviour.

Design motivations

In this section, we describe the key design motivations that users considered important when designing cards which we identified during thematic analysis. Taken together, these motivations help us understand the creative task of card design for *Hearthstone*, and may help shape the design of future co-creative systems in that space.

Class-themed Design: Keeping the cards within the theme of the respective classes: considering class-specific abilities, and designing cards that fit within existing archetypes associated with each class.

Synergy-themed Design: Designing cards that work well together. Designers consider how cards might combine together to be more powerful than they might be individually.

Role-themed Design: Designing cards to serve a clear role in a deck. Users described their cards as being either aggressive or defensive, or designed for early- or late-game play.

Balance and Experience: The desire for cards to perform fairly and lead to a fun user experience. Users describe the kind of impact they want their cards to have on a game, adjusting power levels accordingly.

Flavour and Lore: The non-gameplay aspects of cards such as the artwork, and background lore. One user with extensive experience with *World of Warcraft*, a game from the same fictional universe as *Hearthstone*, designed many cards with their favourite characters as inspiration.

Simulation results

We categorised users' responses to the simulation results, focusing on whether each card's performance (i.e. effect on winrate) and behaviour (i.e. other effects) were expected or surprising. 8 users each designed 9 cards, for a total of 72 card designs in this analysis. In terms of winrate, 34 (47%) of the cards performed as expected, while 38 (53%) were described as "surprising". Behaviour was more predictable, with 56 (78%) cards behaving as expected, and only 16 (22%) simulation results being surprising. In other words, user expectations of how a card would act on the game were relatively accurate, but their understanding of how that card would affect the winrate was no better than random chance. This effect may partially derive from the difference between simulated and actual play, this is significant supporting evidence of the utility of our approach for card design. See Table 1 for the matrix of performance and behavioural surprise.

	Performance expected	*Performance surprising*
Behaviour expected	29	27
Behaviour surprising	5	11

Table 1: Card performance surprise and behavioural surprise.

We also categorised modifications users made after receiving feedback on their cards: minor modifications (i.e. changing mana cost, attack, and health by a few points), significant modifications (i.e. modifying or adding a new effect, or large changes to the card's mana cost, attack, and health), or no modification. Out of 72 cards, users modified 29 (40%) in a minor way, 3 (4%) were significantly modified, and the remaining 40 (56%) cards were not modified. See Table 2 for the matrix of card performance surprise and modification choice. Unsurprisingly, users more frequently modified cards whose performance was surprising. This supports the notion that CardLab can drive design iteration.

	Performance expected	*Performance surprising*
Some modification	10	22
No modification	24	16

Table 2: Card performance surprise and modification choice.

Discussion

We found that our user interface enabled users to create and test basic cards successfully. Users described a wide range of motivations behind their designs, considering how their cards would fit into decks, archetypes, and classes. Users also considered the impact cards would have on gameplay, aiming for balanced, original, and fun designs. Our AI simulation-based performance feedback helped users identify how cards could be redesigned to better achieve their intended impact on games. Given the known centrality of testing and iteration on creative design, we posit that this suggests a strong utility for this kind of "simulation-based" co-creative design system. In addition to potentially being of use in the context of *Hearthstone* card design, this suggests that the co-creative card design task may be an interesting area for future computational (co-)creativity research.

Our analysis of design intent highlighted the variety of design motivations that our participants considered when creating their cards including the potential impact of the cards on games, the player experience, and the health of the overall metagame. However, we also identified motivations which may exist in tension with the desire for originality and balance, such as the desire for cards to match existing archetypes or fit with flavourful ideas. This demonstrates the complexity of the card design task, but also potentially illustrates some directions that future co-creative systems in this space might be able to pursue.

One potential limitation of our study is that some fraction of the users' surprise at the performance of their cards may have been due to the comparatively small number of decks which we simulated. Some users designed quite complex cards that would be impactful only in niche circumstances, such as in combination with two or more other cards, or in specific deck archetypes. It is unlikely that the relatively simple player AI and card-substitution system we used in this study would showcase the strengths of such a card. Nevertheless, a significant portion of user surprise at the performance of their proposed designs appears to have arisen from a genuine expectation mismatch caused by the inherent complexity of balancing a new card in a game like *Hearthstone*. This kind of performance feedback often caused users to reconsider their card design.

Behaviour was more predictable, with many cards resulting in obvious changes to the overall behaviour of a deck (e.g. healing minions leading to more healing done). However, users described behavioural feedback as valuable, helping them better understand the impact their cards would have on games. When surprise was elicited by the behaviour (rather than performance) of proposed cards in our simulated games, it tended to initially exhibit confusion, since the changes were often indirect or secondary impacts of the proposed change. While this kind of surprise was relatively rare in our study (compared to unexpected performance), they indicate moments where the system was able to highlight complex downstream consequences the user might not otherwise have spotted. These surprises led to significant verbal reflection, as well as occasional substantial modifications. We interpret these early signs of reformulation as preliminary evidence of CardLab's capacity to facilitate co-creativity through automated playtesting.

Users sometimes described cards that they wanted to create, but could not due to limitations of our prototype. For example, many users desired more control over summoned minions, such as being able to make a card that summons a particular creature type (e.g. "Battlecry: Summon a 1/1 *Murloc*"). Other users identified a desire to have more control over the specific targeting of effects (e.g. "Destroy all *damaged* minions"), or the ability to invert a selection (e.g. "Destroy all *non*-beast minions"). The simulation engine used in our study would be able to incorporate these effects with ease, the only requirement would be for a more complex card creation user interface.

Overall, however, users found reflecting on the simulation data engaging and useful to their design process. Performance feedback allowed our users to get a better understanding of how their cards could be possibly balanced while behavioural feedback facilitated a greater understanding of card impact. While the CardLab prototype is just that – an initial exploration of the possibility of simulation-based automated playtesting – we believe it has shown the promise of this approach to co-creative game content design.

Future work to develop CardLab's capabilities could explore the system generating original cards or suggested changes to proposed designs automatically. By scaling up the simulations using high-performance computing, it would be possible to evaluate a large range of computer-designed cards, which by implementing quality diversity algorithms could be diverse, balanced, and behave as intended by designers. We also believe that future systems should explore the deck-level and meta-level considerations of card design, factoring in the complex social dynamics which drive metagame lifecycles. Recent developments in image-generating AIs and large language models have also opened up new avenues to explore the automatic creation of non-gameplay elements of cards, such as artwork, lore, and flavour-text. Future systems may be able to design all aspects of a complete card-set and this represents many exciting research directions.

References

Adams, E. 2014. *Fundamentals of Game Design*.

Carter, M.; Gibbs, M.; and Harrop, M. 2012. Metagames, paragames & orthogames: A new vocabulary. In *Proc. of the Int. Conf. on the Foundations of Digital Games*.

Charity, M.; Khalifa, A.; and Togelius, J. 2020. Baba is y'all: Collaborative mixed-initiative level design. In *IEEE Conference on Games*.

de Mesentier Silva, F.; Canaan, R.; Lee, S.; Fontaine, M.;

Togelius, J.; and Hoover, A. 2019. Evolving the hearthstone meta. In *IEEE Conference on Games*.

Dockhorn, A., and Mostaghim, S. 2019. Introducing the hearthstone-ai competition.

Dubois, L.-E., and Weststar, J. 2022. Games-as-a-service: Conflicted identities on the new front-line of video game development. *New Media & Society* 24.

Hoover, A. K.; Togelius, J.; Lee, S.; and de Mesentier Silva, F. 2020. The many ai challenges of hearthstone. *Künstliche Intelligenz* 34.

Liapis, A.; Yannakakis, G. N.; and Togelius, J. 2013. Sentient sketchbook: Computer-assisted game level authoring. In *Proc. of the Int. Conf. on Foundations of Digital Games*.

Smith, G.; Whitehead, J.; and Mateas, M. 2010. Tanagra: A mixed-initiative level design tool. In *Proc. of the Int. Conf. on the Foundations of Digital Games*.

Świechowski, M.; Tajmajer, T.; and Janusz, A. 2018. Improving hearthstone ai by combining mcts and supervised learning algorithms. In *IEEE Conf. on Computational Intelligence and Games*.

Zhang, Y.; Fontaine, M.; Hoover, A.; and Nikolaidis, S. 2022. Deep surrogate assisted map-elites for automated hearthstone deckbuilding. In *Proc. of the Genetic and Evolutionary Computation Conference*.