Replay Enactments: Exploring Possible Futures through Historical Data

Kenneth Holstein *
Carnegie Mellon University
Pittsburgh, PA
kjholste@cs.cmu.edu

Erik Harpstead *
Carnegie Mellon University
Pittsburgh, PA
harpstead@cmu.edu

Rebecca Gulotta
Carnegie Mellon University
Pittsburgh, PA
beka@google.com

Jodi Forlizzi
Carnegie Mellon University
Pittsburgh, PA
forlizzi@cs.cmu.edu

ABSTRACT
As we design increasingly complex systems, we run up against fundamental limitations of human imagination. To support practice, it becomes essential to use authentic data and algorithms as design materials to augment designers' intuitions. Recent work has explored some dimensions of using data as a design material, suggesting the contours of a new space of design and prototyping methods. In this paper, we present *Replay Enactments (REs)*, an extension of the User Enactments methods that uses data replay as a boundary object, making complex system behavior tangible to designers and stakeholders. We reflect on a set of case studies that have instantiated REs in diverse ways and discuss trade-offs between different ways of using data replays in design. We conclude by highlighting opportunities and challenges for future work.

Author Keywords
Replay Enactments, Design Methods, Prototyping, User Enactments, Data Replay, User Experience Design

CSS Concepts
•Human-centered computing~Human computer interaction (HCI)~HCI design and evaluation methods

INTRODUCTION
As we design increasingly complex systems, design teams run up against fundamental limitations of human imagination. Design teams must envision how adaptive systems might behave across a *wide range of possible user interactions and contexts* [8, 18, 36]. Particularly when these systems are targeted for global deployment, failures of imagination often have unintended consequences, for example where the value provided by adaptive services degrades across different groups of users and contexts [18, 30, 33, 38]. When designing data-driven AI systems, design teams must also envision the impacts of *imperfect algorithms*, for example by anticipating how users will experience the errors a particular algorithm makes when fed data from a particular context [8, 17, 18, 35, 36]. Yet such imperfections are challenging, if not impossible, for designers to accurately imagine or approximate through simple simulations [8, 35, 36]. How might design teams more successfully work with materials that defy their capacities for projecting into possible futures?

User Enactments (UEs) are a set of design methods that help teams conduct a *fieldwork of the future* [26]: In UEs, designers construct simulations of possible futures (e.g., through immersive physical sets, lo-fi prototypes, and Wizard of Oz methods), and invite users to participate in multiple enactments of loosely scripted scenarios within these contexts [6, 26, 37]. By emphasizing the co-enactment of multiple scenarios, each representing an alternative vision for the future, UEs can function as a boundary object [29] enabling multidisciplinary teams and other stakeholders to explore uncharted design spaces together [26]. In UEs designers use Wizard of Oz (WoZ) methods and role-playing to enact the behavior of novel technologies or to simulate future social contexts. Yet designers often struggle to imagine the behavior of complex algorithmic systems before they are actually deployed [3, 8, 17, 18]. In the wild, a system's behavior can depend heavily on interactions between particular *data-generating contexts* (e.g., specific socio-cultural settings where a system may be used) and particular *algorithms* (e.g., specific AI models trained on specific datasets that encode specific biases).

To aid designers in bridging this gap, we present *Replay Enactments (REs)*, an extension of the User Enactments methods that uses *data replays* [12, 17, 23, 24] to make complex system behavior tangible to designers and stakeholders. Like UEs, REs allow people to sample multiple possible futures via brief enactments of scenarios within a staged or simulated context. However, in REs designers

* Co-first authors contributed equally to this paper.

DIS '20, July 6–10, 2020, Eindhoven, Netherlands

ACM ISBN 978-1-4503-6974-9/20/07…$15.00
https://doi.org/10.1145/3357236.3395427

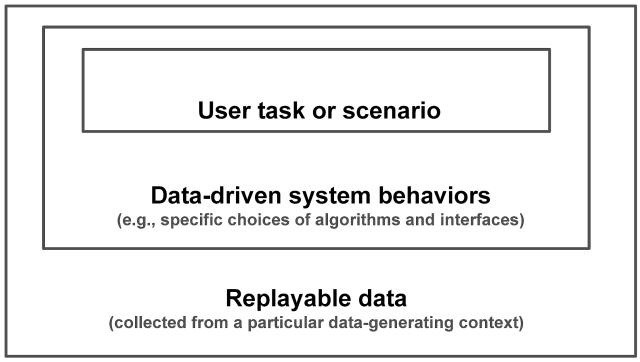

Figure 1. Nested components of a Replay Enactments study, each of which can be swapped out during or between sessions to explore different combinations of data-generating contexts, system behaviors, and scenarios (figure adapted from [17]).

construct these simulations via replays of previously collected data from the field (e.g., metadata from social media profiles or log data from classroom use of educational technologies). These data replays may be enacted either by *human wizards* who interpret and react to incoming data as it is replayed; by *low-fidelity algorithms* [35, 36] that approximate the way an eventual system might behave; or by *high-fidelity algorithms* [14, 17] representing the actual material with which designers will shape a system's algorithmic behavior [1, 3, 36]. The nested components of a RE study, illustrated in Figure 1, can be swapped out during a session to explore different combinations of data-generating contexts, system behaviors, and user tasks or scenarios. Thus for example, across multiple enactments, designers may explore the *same* user scenario using data collected from different contexts, or using different choices of algorithms.

In this paper, we first situate REs in relation to existing design research methods and discuss their use of data replays as a material. We then present three case studies from our own research, illustrating diverse uses of REs: the *Retrospect, Lumilo,* and *RumbleBlocks* projects, each of which is previewed briefly below.

The *Retrospect* project explored potential futures where a system might help people manage, curate, and reflect on the digital information generated over the course of their lives. Designers observed participants' responses to various, alternative re-enactments of participants' own data (scraped from their social media accounts). Here, the use of authentic personal data enabled participants to reason from their own experience in evaluating ways in which their data might be understood and used in the future, while in turn enabling designers to observe nuances that they may not have otherwise imagined into fictional data.

The *Lumilo* project engaged K-12 teachers in iteratively shaping the behavior of data-driven AI systems for use in their classrooms. Here, the goal was to make the experience of working with imperfect algorithms tangible to stakeholders, via immersive simulations of the way these algorithms would behave in different classroom contexts. During a study, teachers experienced a simulated class session based on replays of classroom field data, filtered through the kinds of data-driven algorithms that might eventually be used in a deployed system. The use of authentic data and algorithms provided designers early insight into how teachers' experiences were shaped by the interplay of particular algorithms and classroom contexts. Meanwhile, these simulations enabled teachers to experiment with different algorithmic design decisions and experience the consequences of these decisions.

Finally, the *RumbleBlocks* project helped designers both *retrospectively* understand players' experience in an educational game and *projectively* explore the implications of particular game design alterations on the player experience. Unlike the *Retrospect* and *Lumilo* case studies, where human participants engaged in co-enacting the scenarios, here the enactments were fully machine- performed. In the *retrospective* case, historical gameplay data were re-enacted to materialize the dynamics of an educational game and allow designers to examine the same sessions from multiple analytical perspectives. In the *projective* case, historical gameplay data was used to simulate the impacts of potential design changes, to help designers evaluate whether the resulting state of the game would be preferred to the current one.

Drawing upon these case studies, we conclude with reflections on how and when to use REs, highlighting opportunities and challenges for future work.

REPLAY AS A META-MATERIAL IN DESIGN

A core contribution of this work is the positioning of data replays as a useful *meta-material* in the design process. By this we mean that data replays, as concrete abstractions of specific user experience and context, can be re-enacted to provide designers a lens on a potential future user experience. They are a meta-material in the sense that the data replay is not the material object of the design process directly but rather the *means to materializing* the real object of design. By manipulating what aspects of data are being replayed and how the replay is being executed and interpreted, the designer can explore the material implications of design choices.

Across our three cases, we have used data replays to support each phase of an iterative design process, from *observation* to *ideation* to *iteration* [6]. The use of replays supports both *observation* and *ideation* by allowing designers and participants to experience data collected from the field in *new ways*. Through multiple enactments of the same interaction traces, data replays enable multiple re- experiences and re-interpretations of the same set of field observations [10, 14], using different abstractions to make different aspects of an experience more or less salient [12, 16]. In addition, the use of data replays supports *ideation* and *iteration* by making complex system behaviors tangible.

	Authentic field data	Staged / simulated field context	Co-enactment by humans + machines	Multiple brief enactments
User Enactments		●		●
Speculative Enactments		●		
Field Deployments	●		●	
Replay Enactments	●	●	●	●

Table 1. Comparison of Replay Enactments and closely related approaches (rows), along four defining properties (columns).

Through REs, designers and other stakeholders can rapidly, iteratively explore the UX implications of otherwise opaque design decisions (e.g., choices of data, AI models, or parameter settings) by playing out the impacts of a change and materializing the results [17, 36].

Table 1 presents four defining properties of REs, contrasting against related approaches including technology field deployments (e.g., [20, 25]) and prior enactment- based approaches that help teams explore potential futures in staged environments. In the latter category are the WoZ-based User Enactments (UEs) approach, described above, and the improvisation-based Speculative Enactments (SEs) approach [9]. Drawing upon UEs and speculative design, SEs is an approach aimed at engaging study participants in speculative yet personally consequential circumstances. Unlike UEs, SEs minimize the use of WoZ methods or scripting of interactions, instead prioritizing the creation of conditions for genuine social interactions to unfold among participants. As such, SEs rely more heavily on participant *improvisation* within the broad premises of a given speculative future. Unlike UEs and REs, which emphasize the exploration of many possible futures via multiple brief enactments (see *Multiple brief enactments* in Table 1), the SEs approach favors fewer and longer enactments with the goal of fostering deeper participant investment in particular futures [9].

Both SEs and the WoZ-based UEs approach, in which human actors may imagine and enact possible machine behaviors, can be useful in rapidly exploring broad, uncharted design spaces earlier on in the design process. For example, an early UE study [6, 37] explored a diverse range of smart home concepts, embedded in different household scenarios, to help designers anticipate parents' desires and boundaries regarding a smart home's roles within family life. Since both the behavior of a potential system and the surrounding context in UEs and SEs are animated by the designers and participants themselves, they are free to imagine any number of possible interactions, including ones that are not (yet) feasible with existing technologies. However, while these wider potential futures are afforded by imagination, they are also bounded by it.

WoZ or improvisation-based approaches alone may fail to capture real world dynamics that are critical to a user's experience. For example, human wizards and actors may not accurately imagine the nuances of authentic family interactions and routines in the home [8, 17] (see *Authentic field data* in Table 1). Further, even when WoZ is successful in convincing participants that they are interacting with a machine, human wizards may fail to enact realistic machine behavior, limiting what designers are able to learn from a study. In particular, WoZ can fail to represent the behavioral complexity of data-driven algorithmic systems, such as the patterns of inference errors that an actual smart home might make [8, 17, 36]. As discussed in recent work on *AI as a design material*, when designing complex algorithmic systems, it may become necessary to explore the material properties of *actual* algorithms earlier on in the design process (e.g., [3, 8, 17, 35, 36]). This may be achieved by moving from fully human-performed enactments, as in UEs and SEs, towards approaches where machine behaviors are enacted (at least in part) by actual machines (see *Co-enactment by humans + machines* in Table 1).

While each of these aspects – real world field dynamics and algorithmic behavior – can be explored by developing and deploying technologies in actual field contexts, field deployments can be costly and may limit the extent to which design teams can freely explore and iterate (see *Staged or simulated field context* in Table 1). REs can be used to bridge the gap between UEs or SEs and technology field deployments, reducing risk as designers make the leap from prototyping semi-scripted system behaviors in semi-controlled contexts to prototyping complex system behaviors in messy and diverse field contexts. For example, in the *Retrospect* case, the use of participants' personal data within the context of a RE study helped users of the system to reflect on personal experience without feeling as if their privacy had been violated.

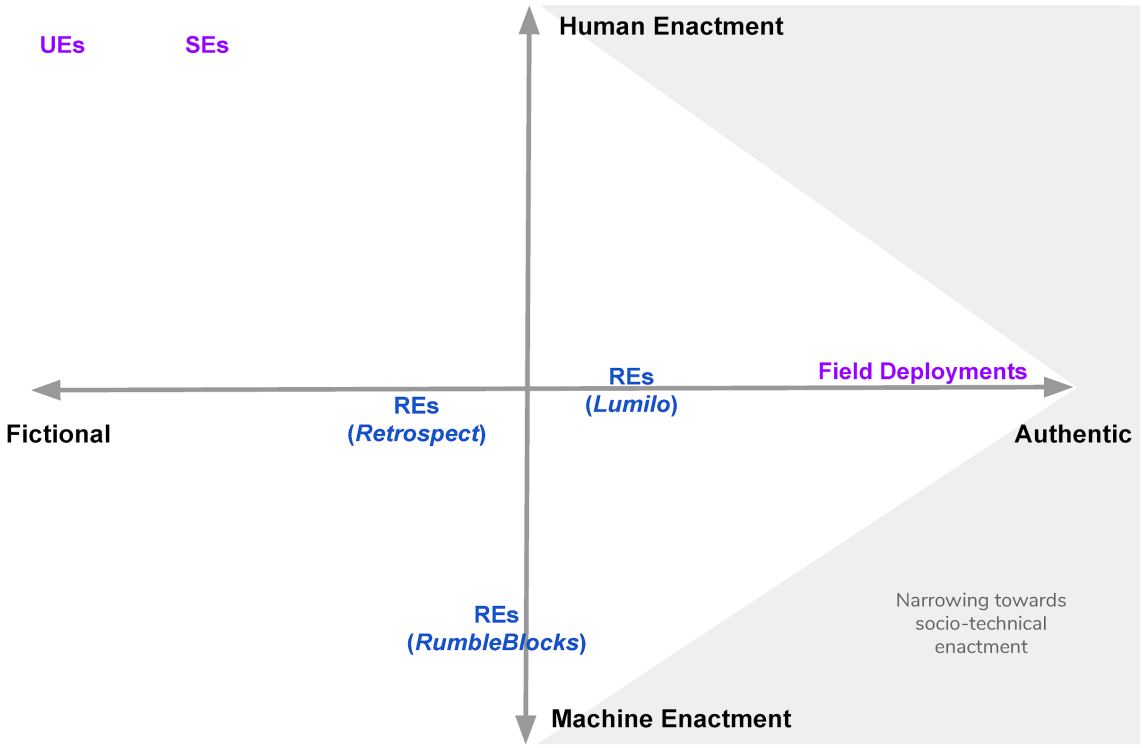

Figure 2. A continuum of methods for exploring possible futures, in two dimensions: The horizontal axis represents the extent to which the task of anticipating designed systems' behavior and real-world impacts is informed by actual field data and algorithmic behavior (right), versus being imagined (left). The vertical axis represents the extent to which scenarios are enacted by human participants (top) or by machines (bottom). As design teams converge towards technology field deployments, scenarios increasingly unfold through an interplay of human and machine behavior. This is illustrated by a narrowing of the possibility space moving towards the right of the diagram (grey boundaries) towards *socio-technical* enactment: real humans interacting with real technologies in real social contexts. Replay Enactments (REs) occupy a middle region of this space, between purely WoZ- or improvisation-based approaches and technology field deployments. Shown in blue: The three case studies presented in this work.

While this differs from a real-world study, designers were able to weigh the benefits and tradeoffs of combining personal data that were otherwise siloed over a number of sources. Similarly, in both the *Lumilo* and *RumbleBlocks* cases, a key motivation for using REs was to anticipate the real-world impacts and behavior of complex algorithmic systems, where moving straight ahead to a field deployment was viewed as too risky. Even after a field study, REs can be used to enable continued iteration outside the constraints of the field, as in the *RumbleBlocks* case.

By reflecting on diverse instantiations of Replay Enactments in this paper (cf. [17]), we intend to open up a middle space of prototyping methods, between purely WoZ- or improvisation-based approaches and technology field deployments. Both REs and technology field deployments tend to require greater upfront technical investment than UEs or SEs. However, as with technology probes [20], artefacts developed for REs are primarily intended as tools for design exploration, rather than as fully-fledged prototypes. As shown in Figure 2, specific instantiations of the REs approach can span a range of locations within this middle space. In navigating this space, design teams face trade-offs between flexibility and realism.

As design teams move further towards the right of the space, prioritizing authenticity, they are better able to ground their speculations about possible futures in the dynamics of particular real-world contexts and the limitations of actual algorithms. At the same time, moving further towards the right side may increasingly anchor and limit the kinds of futures that can be envisioned to those contexts and algorithmic capabilities that *exist now* – potentially at the cost of envisioning futures that diverge more radically from the present.

REs towards the left in Figure 2, prioritizing flexibility, may involve enacting scenarios based on authentic data from the field (e.g., previously collected interaction traces) but without necessarily using authentic algorithms. Such REs may be enacted either by human wizards who interpret and react to replayed data (see *Discussion*), or by "low-fidelity" versions of the kinds of algorithms that may eventually be fielded (e.g., simple rule-based simulators intended to approximate the behavior of machine learning systems [36]). When it is important to prototype realistic algorithmic behavior (as in the *Lumilo* case), REs towards the right of Figure 2, using *both* authentic data and algorithms, may be most informative. Similarly, as design teams converge towards deploying technologies in actual field contexts – where scenarios will

evolve through an interplay of human and machine behavior (see grey boundaries in Figure 2) – it may be increasingly important for REs to engage human participants in *co-enacting* scenarios [17] with machines (as in the *Retrospect* and *Lumilo* cases).

In positioning data replays as a meta-material, we see Replay Enactments as supporting an expansion of the craft orientation to HCI design process [21, 34]. Rather than moving away from a designerly approach, introducing authentic data and algorithms offers designers the ability to reify user experience into a material form that they can see and converse with [8, 30, 35, 36].

CASE STUDIES

In the following sections, we reflect upon experiences conducting REs across three of our own projects, which have instantiated the method in diverse ways.

Retrospect: Building User Enactments from Metadata

In this project, Gulotta et al. expanded upon the User Enactments methods to create a demonstrational prototype called *Retrospect* [10]. *Retrospect* was designed to reflect potential futures where a system might help make sense of, manage, and represent the digital information generated over the course of one's life (Figure 3). Examining these issues is challenging because few systems can gather or analyze information on this scale or for this purpose.

Digital systems capture an increasingly large and significant portion of people's life experiences. It is important to consider how people navigate the processes of managing, curating, and reflecting on that information. The field of personal information management has attempted to develop systems and practices to help people better manage and locate pieces of digital information. However, personal digital information is idiosyncratic and fragmented across identities and services. The *Retrospect* project expanded upon efforts in personal information management to explore how people might manage, curate and archive records that span across lifetimes and generations.

Enactments

Retrospect relied on metadata about participants, scraped from social media. We explored two categories of metadata: (1) person-generated metadata, such as comments on a Facebook post, and (2) system-generated metadata, such as the number of times a song has been played. Metadata is one of the main sources of information that systems capture about users and leverage to make decisions about what information to share with those users. However, the degree to which users are aware of having contributed this data greatly influences how they perceive system actions.

The goal in this study was to scrape participants' personal data to use in familiar scenarios, to help participants reason from their own experience in evaluating *Retrospect*. After the initial set up, *Retrospect* was used by participants for two and a half months. Each week, participants engaged in reflective tasks prompted by *Retrospect*. Across a series of interviews that

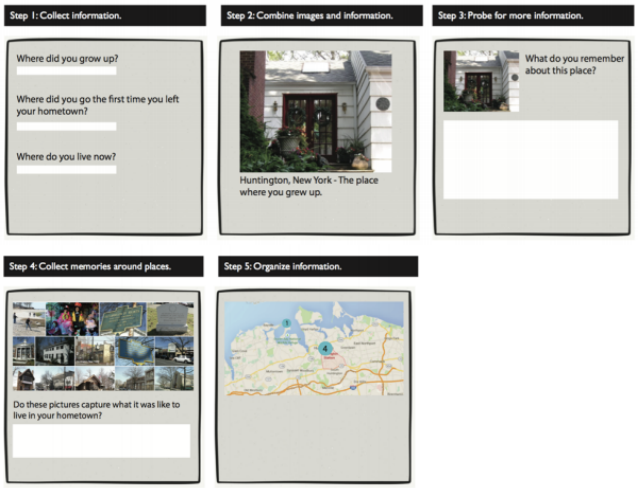

Figure 3. Illustration of one potential design for *Retrospect*, which combined personal data across multiple sources, using locations from a person's past as a resource for reflection [11].

took place across the duration of the study, designers observed participants' responses to alternative ways of re-enacting participants' own data.

The use of authentic personal data enabled participants to more sensitively consider how users understand the quality and texture of the many different types of digital information to which they are connected. Additionally, variations in the data (e.g. when it was originally created, what kind of information it was, what part of one's life the data reflected), prompted participants to imagine how future systems might make use of the wide range of data about their lives, behaviors, and experiences. In turn, this enabled designers to observe nuances which they may not otherwise have imagined into fictional data.

Though a number of data sources were considered, the *Retrospect* study utilized a participant's Facebook's data. Facebook's API functionality was well documented and there was a large group of developers using the API. The popularity of Facebook helped ensure that adequate data could be scraped, and that *Retrospect* would have access to data in both the near and distant past. This decision resulted in constraints for the design of the *Retrospect* system [10, 11] but served as an entry way for participants to reflect on the great variety of data available about them online.

Reflections

Conducting REs with participants' personal data enabled designers to observe participants' responses to various, alternative re-enactments of this data. A key aspect of *Retrospect* was that it prompted users to reflect on pieces of digital media and information from different parts of their own lives. Interviews with participants revealed that few of the participants engaged in deliberate, unprompted revisitation of their digital content without *Retrospect*. Participants liked how *Retrospect* eliminated the step of hunting for old content on a variety of volumes such as external hard drives or obsolete

phones. Many of the participants enjoyed having the time and motivation to revisit aspects of their past. Interestingly, no one expressed negative sentiment about the use of disparate bodies of personal information to create the *Retrospect* system. The design team found that using participants' personal data in this way allowed for deeper personal reflection. The addition of design features beyond those expressed through the scraped data helped to create a personalized and meaningful experience.

While these REs were constructed from authentic historical data, exploring the UX of authentic data-driven algorithms was not a major goal of this study. Thus, the *Retrospect* project's REs were slightly further from a technology field deployment, in terms of technical realism, than the cases we present next (see Figure 2's horizontal axis). Along the vertical axis in Figure 2, the *Retrospect* study falls near the center: while the prototype re-enacted participants' personal data, participants played active roles as users of the system.

***Lumilo*: Co-shaping data-driven algorithmic behavior**

Our second case study involves the use of REs to engage K-12 teachers in iteratively shaping the behavior of data-driven AI systems [15, 17]. Here, the goal was to make the experience of working with imperfect algorithms tangible to teachers, via immersive simulations of the way these algorithms would behave in specific contexts. Thus, it was important not only to use authentic data in these REs, but to enact replays of this data through the kinds of data-driven algorithms that might be used in an actual fielded system.

Holstein et al. used REs to iteratively prototype a real-time decision-support tool for K-12 teachers called *Lumilo* [16]. *Lumilo* is a set of mixed reality smart glasses designed for use in self-paced classrooms where students work with AI-based tutoring systems [4, 28]. As a teacher walks throughout the room while wearing *Lumilo*, they can see real-time indicators about students' learning, metacognitive, and behavioral states, floating directly above students' heads [2, 17]. The underlying constructs behind these real-time indicators (e.g., whether a struggling student is facing productive difficulties versus genuine roadblocks) were selected through a co-design process with teachers, with the goal of alerting teachers to unfolding classroom situations that may benefit from human intervention. Many different algorithmic approaches have been proposed in the literature to measure each of these constructs [7, 17]. However, the "matchmaking" process between algorithms and teacher needs was far from straightforward, and the use of these algorithms for teacher decision-support was an uncharted design space at the start of the project.

Enactments

Fielding prototype systems in K-12 settings risks causing harm to students if the prototype's effects are poorly understood. To rapidly prototype the experience of using *Lumilo* prior to a risky pilot in actual classrooms, the design team made use of data replays, based on previously recorded software log data

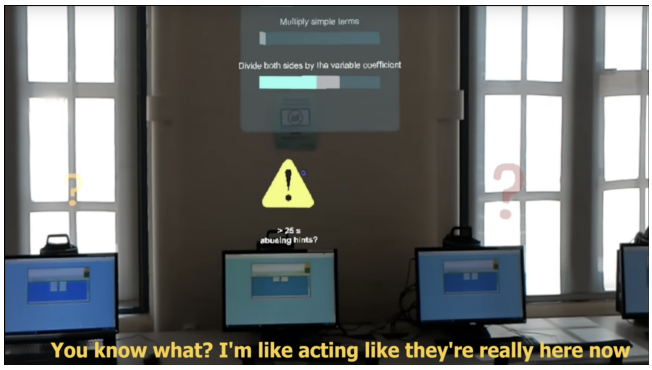

Figure 4. Screenshot showing a teacher's point of view during a RE study with *Lumilo*. Logged interaction data from a full class of students is replayed, at original speed, through AI tutor interfaces on separate computers in the lab; corresponding indicators update in real time through the glasses. Teacher dialogue is displayed at the bottom: in the midst of an enactment, this teacher notices they have begun talking to students as if they were actually present.

from classrooms working with AI tutoring systems. These data were available in LearnSphere, a major repository of educational data, intended for use by researchers and data scientists in conducting offline quantitative analyses [19, 27]. In this case, the team appropriated these existing classroom data for use in Replay Enactments.

The designers brought middle school math teachers into computer labs on their university's campus to participate in RE sessions. In each study session, the teacher wore the *Lumilo* smart glasses while a 40-minute class session was replayed from beginning to end, at actual speed. On each computer monitor in the lab, a different student's actions were replayed through the AI tutoring software. Through the *Lumilo* prototype, teachers saw mixed reality indicators floating above each empty seat, updating in real time.

Teachers were asked to pretend that this was an actual class session, role playing and thinking aloud while they moved throughout the lab space (see Figure 4). If a teacher thought they might focus their attention on a particular student at a particular moment, based on the information they saw, they were instructed to verbalize what they might say to that student if the student were actually there [15, 17].

When iteratively prototyping *Lumilo* with teachers, the same iteration of *Lumilo*'s design was frequently tested across replays of datasets from a range of classroom contexts (e.g., remedial versus gifted classrooms) while holding the choice of algorithms constant, or vice-versa (see Figure 1). During a session, a human "wizard" would make live changes to algorithmic elements of a system's design based on stakeholder feedback. In an iterative fashion, the wizard would elicit design feedback, make small adjustments (e.g., tweaking parameters in a machine- learned model), and allow stakeholders to experience the consequences of their requested changes.

Reflections

Conducting REs with authentic data and algorithms enabled insights into how participants' subjective experiences are shaped by the interplay of particular algorithms (e.g., specific machine learning models trained on specific student datasets) and data-generating contexts (specific kinds of classroom environments from which data were collected). For instance, prototyping sessions with *Lumilo* revealed that under particular classroom dynamics, but not others, teachers experienced the patterns of errors made by particular choices of algorithms as anxiety-inducing [15]. In addition, these REs provided early insight, before entering the field, into the effects that different algorithm design choices might have on teachers' *behavior*. In this case study, the use of data replays removed the possibility that teachers' behavior could influence the replayed students (thus removing the possibility of causal feedback loops). Thus, the team was able to collect and analyze data on teacher behavior in RE sessions to investigate *Lumilo*'s effectiveness in steering teachers' attention towards students most in need of help in-the-moment.

As in the *Retrospect* case study, these REs engaged human participants in *co-enacting* scenarios together with technological components (see Figure 2). Given that the user task in this case – monitoring a classroom and guiding students – is highly embodied and interactive, *Lumilo*'s REs emphasized physical enactments and role playing exercises in the spirit of the UE approach [17, 26, 37]. The instantiations of REs we have discussed in the *Retrospect* and *Lumilo* cases both focus on exploring alternative re-enactments of historical interaction data. Our next case study expands beyond this paradigm: in addition to re-enacting past data, the *RumbleBlocks* case leverages *models* of these data to aid designers in imagining other plausible user behaviors and system dynamics.

RumbleBlocks: Projecting Playtests into Alternate Futures

Our third case study explores the use of REs to both *retrospectively* understand the player experience in an educational game and *projectively* explore the player experience implications of changes to the game's design. The case centers on the iterative evaluation and design of *RumbleBlocks* (see Figure 5), an educational game intended to teach young children concepts of structural stability and balance by having them build block towers that have to survive earthquakes [14]. The complex algorithmic design challenge in this case was dealing with the indeterminacy of the game's physics engine, which was used to evaluate players' solutions. Given the educational purpose of the game it was important to ensure an alignment [13] between the feedback students received and the stated principles the game was teaching by tuning properties of the physics system and other mechanics of the game.

In a traditional game design setting this alignment would be achieved through extensive iterative playtesting [5]. Given the

Figure 5. A screenshot of a level from the *RumbleBlocks* game (left) and a visual representation of four different player solutions for this level (right). In the replays, non-essential game elements were hidden, and each free-standing tower was color coded to aid in interpretation.

teducational goals of the game, it was important to the designers that this testing be done with children in the target demographic so as to design against realistic learner misconceptions. However, exploring several iterations of game mechanics with a realistic player population in a realistic classroom setting would place a substantial burden on local schools by organizing many disruptive classroom playtests. Thus it was important for the designers to be able to explore realistic player behaviors and corresponding game responses for a wide range of possible mechanics outside of the field, while still being informed by it.

Enactments

In this project REs were used to augment the value of the limited window of data afforded by rare and costly classroom playtests. The data for the REs was gathered during an initial field trial of *RumbleBlocks* that was done to gauge its educational effectiveness [12, 13]. This field trial involved *in vivo* classroom playtests of the game with 281 students across the games' K-3 (5-8 year-old) target demographic in two local schools. In addition to assessment instruments used for evaluation purposes, each student's gameplay session was recorded as a replay trace that could be simulated in the game engine after the testing session.

REs were used in 2 distinct phases in this work. The first phase was to perform REs retrospectively to help the designers observe the current state of the game's design. In this retrospective phase the replay engine used to re-enact game sessions was instrumented to produce data for various game analytics techniques. For example, calculating metrics from player solutions to measure whether following target principles corresponded to success in the game [14] or abstracting player solutions into build patterns to observe different ways players approached game levels [13]. This allowed the designers to reframe the in-class playtests in multiple ways to consider whether the current design supported their instructional goals. In doing so, they found that the way the game evaluated player solutions to in-game puzzles was often skewed by micro-faults in the solution

rather than capturing holistic physical properties, which was the instructional goal of the game [12].

After discovering faults in the game using retrospective REs, the designers of *RumbleBlocks* considered several potential design solutions to the issues they observed. This led to a second phase of the work where the REs were used projectively to simulate the impacts of potential design changes and consider whether the resulting state of the game would be preferred to the current one. In this context, projective REs leveraged the existing replay data as the basis for an AI player model of how a target user population would react to the current design. They then leverage the affordance of a replay environment to project that model onto an iterated version of the original game to help designers anticipate how players might play the new version differently.

In these projective REs, not only is the product itself simulated but so is the user population. A key commitment of this approach was to rely on a cognitive architecture designed to replicate the human learning process [22], which presents similar imperfections to a human learning something for the first time, rather than a superhuman AI approach [32]. In this way the approach attempts to create a middle space between running authentic, but disruptive, playtests with new users, versus entirely theoretical simulation techniques, uninformed by actual human performance or context.

Reflections

In the work on *RumbleBlocks,* projective REs supported rapid, divergent iteration of the game. In a particularly salient example, the technique was able to catch a subtle problem with one redesign's scoring mechanic where players would fail levels without a clear sense of why their solution was wrong [12]. While these issues might have turned up in playtests with human players, and indeed did in some cases, the designers were able to explore many more design concepts than they would have been able to with classroom playtests alone, given logistical constraints.

Unlike the *Retrospect* and *Lumilo* case studies, where human participants engaged in co-enacting scenarios, here the enactments were fully machine-performed (falling towards the bottom of the vertical axis in Figure 2). The *RumbleBlocks* case is also unique in that, in addition to re-enacting past data directly, designers leveraged generative *models* machine-learned from these data to more flexibly explore possible futures. In this sense, *RumbleBlocks'* REs fall to the left of *Lumilo's* REs along Figure 2's horizontal axis, providing designers greater flexibility at the risk of trading off some realism.

TRADE-OFFS AND LIMITATIONS

The *Retrospect*, *Lumilo*, and *RumbleBlocks* case studies illustrate three points within a broader space of methods that use data replays to support design. Across these cases, Replay Enactments supported design teams in bridging the often large gap between prototyping semi-scripted system behaviors in semi-controlled contexts (e.g., through conventional User Enactments) to prototyping complex system behaviors in messy and diverse field contexts. In the following, we discuss major trade-offs across this space of methods, reflecting on where particular approaches may provide the most value.

Flexibility versus Realism

Design teams face challenging trade-offs as they navigate the methodological space shown in Figure 2. Towards the left end of this space, characterized by WoZ and improvisation-based approaches, designers have the most flexibility. Since both the behavior of a potential system and the surrounding context are animated by the designers themselves, they are free to imagine any number of possible interactions, including ones that are not (yet) feasible with existing technologies. At the same time, while these wider potential futures are afforded by designer imagination, they are also bounded by it. For instance, a designer's capacities for imagination will be limited by their own prior background and experience. Data-driven products and services often encode assumptions reflecting the demographics of the design team, which can have unintended consequences when designing for a global context [18, 30, 33, 38].

REs can help to augment design teams' capacities for imagination, as well as those of other participants who are engaged in these enactments. In cases where a design team does not have immediate or sustained access to particular groups of stakeholders who will use or be affected by a new technology, replays of recorded experiences from these groups can provide a partial window into these stakeholders' needs. For example, through retrospective REs, design teams might use a targeted sample of data from an underrepresented population to sensitize themselves to the experiences of members of that group within their product. Today this type of work is often done using personas, which do not have the same capacity to evoke the complex emotions, behaviors, and insights as re-enacting a person's actual data.

Although REs can augment and extend design teams' imaginative capacities as discussed above, the method's reliance upon existing data and algorithms can be limiting in other ways. As teams allow their enactments of possible futures to be guided by authentic data and algorithms, the kinds of futures they are able to envision may be increasingly anchored to (and limited by) current algorithmic capabilities, user interactions, and field dynamics [8, 12, 17]. The realism and nuance this affords comes at the cost of exploring futures so radically different from the present that assumptions of existing data and algorithms break down – a strength of User Enactments and related methods such as Speculative Enactments. One way of balancing these trade-offs may lie in projective replay approaches, as in the *RumbleBlocks* case, which aid designers in more flexibly generalizing patterns from historical data to new situations. Another approach may be to develop hybrid approaches that combine the flexibility of human wizards and actors with the realism of data replays (see *Future Directions*).

Technical Investment and Constraints

In contrast to conventional UEs, which support design teams in investigating a wide range of alternative concepts early on in the design process, the RE case studies presented above explored comparatively narrower design spaces. One reason for this is that the infrastructure necessary to produce an interactive replay of data requires a level of upfront technical investment that constrains the design space in particular directions.

In each of our three case studies, the team needed to instrument prototypes so that they could be controlled through historical interaction data. The process was constrained by the availability of authentic data capable of representing realistic user behavior, as well as the availability of a palette of existing algorithms that may provide desired functionality. For example, after committing to the use of the Facebook API as a data source, the *Retrospect* RE study was necessarily scoped to exploring alternative ways of re-enacting participants' Facebook data as opposed to other authentic data sources. To work around this issue, interviews with participants used this Facebook data as a starting point for discussions about the other types of data captured by other systems. Similarly, after committing to the use of classroom interaction data from LearnSphere, the *Lumilo* RE study was scoped to exploring a palette of existing student modeling algorithms that were compatible with this data source. Such technical investment and constraints may be undesirable at the earliest stages of a design process. However, as discussed in recent work on data and AI as design materials, when designing complex algorithmic systems, it may be *necessary* to commit to such technical investment slightly earlier in the process, to reduce risk before introducing such systems into real world field contexts (e.g., [3, 8, 17, 35, 36]).

The specific level of technical investment required to conduct an RE depends on a design team's goals, and which aspects of a potential future experience the team wishes to materialize. In some cases, it may suffice to conduct REs with historical interaction data, but without necessarily using the kinds of algorithms with which designers will shape a system's algorithmic behavior [1, 3, 36]. For example, in the *Retrospect* case, the primary goal of using authentic personal data was to help participants reason from their own experiences, and to help designers observe nuances in the enactment and experience of actual data which they may not otherwise have imagined into fictional data. In other cases, a central goal may be to materialize the behavior and impacts of actual (imperfect) algorithms, as in the *Lumilo* and *RumbleBlocks* case studies. In such cases, conducting REs with approximations to the actual design material (e.g., using simple rule-based simulators to approximate statistical machine learning algorithms) can prove insufficient [30, 35, 36].

FUTURE DIRECTIONS

The discussion of methodological trade-offs above suggests a broader space of methods that might productively (1) combine advantages of WoZ approaches with the advantages of using authentic data and/or algorithms, or (2) combine advantages of field studies with advantages of REs. Possibilities within each of these directions are briefly discussed below, spanning less-explored regions of the methodological space shown in Figure 2.

Exploring Human–Replay Hybrid Approaches

New hybrid approaches might engage human wizards in using replays to *inform* their own enactments, while still retaining the ability to flexibly improvise beyond these data. For example, a human wizard might monitor data replays in real-time, using authentic data and algorithmic behavior to inform manual enactments of system behaviors. Alternatively, rather than having replays play a backstage role, human–replay hybrid approaches may involve *collaborative enactments* between human wizards and replay agents. For instance, an extension of the approach taken in the *Lumilo* case study might occupy some seats in the computer lab by human actors (playing the role of students), while populating the rest of the class with replays of historical student interactions. Compared with a fully replay-based approach, this combination of multiple human and replayed participants in an enactment may enable design teams to more effectively explore social interactions among multiple participants (a strength of User Enactments and Speculative Enactments). At the same time, the use of replayed field data may capture nuances in the dynamics of real-world classrooms, which may not be reproduced by actors placed in an artificial environment (e.g., patterns across multiple students' interactions in real classrooms that arise due to genuine social ties among students [16, 25]).

Combining both human and replayed participants may enable design teams to conduct User Enactments with larger *systems of users* than would otherwise be practical (e.g., classrooms of 20-40 students, where it may be impractical to bring all participants into the lab). As another example of an approach in this methodological space, multiple human wizards may work collaboratively to enact realistic behavior of *individual components* of a complex algorithmic system, potentially in collaboration with actual algorithmic components. For example, Yang et al. [35] involved multiple human wizards in enacting the roles of different components in a generative neural network. Each wizard would generate specific kinds of outputs, which were then combined via an algorithm that assigned different weights to each wizard. In doing so, the system of wizards was able to simulate different kinds of realistic output errors. Further exploring how WoZ methods might be combined with authentic data and algorithms represents a fruitful direction for future work.

Extending Fieldwork through Replay Enactments

Another promising direction involves combining the affordances of REs with those of field studies. For example, a design team might re-enact trace data from a field study, conducting REs with participants from the original study shortly following the study's conclusion. In addition to

supporting retrospective contextual inquiries in cases where interruptions to live field scenarios would be undesirable, these field "follow-up" REs could support designers and participants in envisioning alternative futures (while important contextual details from their original field experience remain fresh in memory) [15, 16].

REs may also be used to provide a richer window into experiences that would be challenging or infeasible to study in live field contexts. For example, in cases where running multiple field studies would be costly or ethically fraught, design teams may instead explore multiple alternatives via re-enactments of the same set of field data. Similarly, where observations of a particular user group are rare, projective forms of REs (as in the *RumbleBlocks* case) could be used to amplify the experiences of the population – simulating how a system might behave for similar users even if none have ever had a particular experience.

Finally, a further direction for future work is to stretch the value of field data by exploring the material properties of replays themselves. While we have thus far considered replays as abstractions of whole user experiences, there is no requirement for them to maintain their wholeness. As data recordings, they can be just as freely manipulated and remixed as the systems they are recorded in. This affordance of replay could enable designers to explore the potentials of *interleaved experiences*: stitching together segments of recorded experience across multiple users or contexts to construct new trajectories that are nonetheless grounded in real experiences. Similar to projective replays, such an approach could enable designers to further probe the outer edges of a design space and interrogate the limits of the possible, while still leveraging historical data to augment what they are able to imagine.

CONCLUSIONS

In this paper we have presented Replay Enactments as an extension of the User Enactments methods. These techniques leverage data replays as a meta-material in the design process, to provide a window into possible future user experiences with complex algorithmic systems. The use of data replays can support designers and other stakeholders in playing out the implications of particular algorithmic design decisions across diverse potential field contexts. Collectively, we view this work as mapping out the contours of a "middle space" of prototyping methods – between purely WoZ- or improvisation-based approaches and technology field deployments – that use authentic data and algorithms to support design craft in an increasingly complex and connected landscape.

ACKNOWLEDGMENTS

This work was supported in part by IES grants R305B090023 and R305B150008 to CMU. The opinions expressed are those of the authors and do not represent the views of IES. Special thanks to all study participants and our anonymous reviewers, as well as Gena Hong, Peter Schaldenbrand, Mera Tegene, Mary Beth Kery, Michael Madaio, Vincent Aleven, Bruce McLaren, Jonathan Sewall, and Octav Popescu.

REFERENCES

[1] Oscar Alvarado and Annika Waern. 2018. Towards Algorithmic Experience. *In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, 1–12. https://doi.org/10.1145/3173574.3173860

[2] Pengcheng An, Kenneth Holstein, Bernice d'Anjou, Berry Eggen, and Saskia Bakker. 2020. The TA Framework: Designing real-time teaching augmentation for K-12 classrooms. To appear in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*.

[3] Eric PS Baumer. 2017. Toward human-centered algorithm design. *Big Data & Society* 4, 2: 205395171771885. https://doi.org/10.1177/2053951717718854

[4] Benedict du Boulay. 2019. Escape from the Skinner Box: The case for contemporary intelligent learning environments. *British Journal of Educational Technology (BJET)* 50, 6: 2902–2919. https://doi.org/10.1111/bjet.12860

[5] Judeth Oden Choi, Jodi Forlizzi, Michael Christel, Rachel Moeller, Mackenzie Bates, and Jessica Hammer. 2016. Playtesting with a Purpose. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '16),* 254–265. https://doi.org/10.1145/2967934.2968103

[6] Scott Davidoff, Min Kyung Lee, Anind K. Dey, and John Zimmerman. 2007. Rapidly Exploring Application Design Through Speed Dating. In *Proceedings of the 2007 International Conference on Ubiquitous Computing (UbiComp '07)*. Springer Berlin Heidelberg, 429–446. https://doi.org/10.1007/978-3-540-74853-3_25

[7] Michel C. Desmarais and Ryan S. J. d. Baker. 2012. A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction (UMUAI)*, 22, 1–2: 9–38. https://doi.org/10.1007/s11257-011-9106-8

[8] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, 278–288. https://doi.org/10.1145/3025453.3025739

[9] Chris Elsden, David Chatting, Abigail C. Durrant, Andrew Garbett, Bettina Nissen, John Vines, and David S. Kirk. 2017. On Speculative Enactments. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17).* ACM, 5386–5399. https://doi.org/10.1145/3025453.3025503

[10] Rebecca Gulotta. 2018. Digital Systems and the Material of Legacy: Supporting meaningful interactions with

multigenerational data. *Unpublished doctoral dissertation, Carnegie Mellon University.*

[11] Rebecca Gulotta, Aisling Kelliher, and Jodi Forlizzi. 2017. Digital systems and the experience of legacy. In *Proceedings of the 2017 Conference on Designing Interactive Systems (DIS '17)*. ACM, 663-674.

[12] Erik Harpstead. 2017. Projective Replay Analysis: A Reflective Approach for Aligning an Educational Game to its Goals. *Unpublished doctoral dissertation, Carnegie Mellon University.*

[13] Erik Harpstead, Christopher J. MacLellan, Vincent Aleven, and Brad A Myers. 2014. Using extracted features to inform alignment-driven design ideas in an educational game. In *Proceedings of the 2014 CHI Conference on Human Factors in Computing Systems (CHI '14)*, ACM, 3329–3338. https://doi.org/10.1145/2556288.2557393

[14] Erik Harpstead, Brad A. Myers, and Vincent Aleven. (2013). In search of learning: Facilitating data analysis in educational games. In Proceedings of the *2017 CHI Conference on Human Factors in Computing Systems (CHI'17).* ACM, 79-88. https://doi.org/10.1145/2470654.2470667

[15] Kenneth Holstein. 2019. Designing Real-time Teacher Augmentation to Combine Strengths of Human and AI Instruction. *Unpublished doctoral dissertation, Carnegie Mellon University.*

[16] Kenneth Holstein, Bruce M. McLaren, and Vincent Aleven. 2017. SPACLE: Investigating learning across virtual and physical spaces using spatial replays. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference (LAK'17).* ACM, 358-367. https://doi.org/10.1145/3027385.3027450

[17] Kenneth Holstein, Bruce M. McLaren, and Vincent Aleven. 2019. Co-designing a real-time classroom orchestration tool to support teacher–AI complementarity. *Journal of Learning Analytics (JLA)*, 6(2), 27-52. https://doi.org/10.18608/jla.2019.62.3

[18] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?.In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19)*. ACM, 1-16. https://doi.org/10.1145/3290605.3300830

[19] Kenneth Holstein, Zac Yu, Jonathan Sewall, Octav Popescu, Bruce M. McLaren and Vincent Aleven. 2018. Opening up an intelligent tutoring system development environment for extensible student modeling. In *Proceedings of the 2018 International Conference on Artificial Intelligence in Education (AIED'18).* Springer, Cham, 169-183. https://doi.org/10.1007/978-3-319-93843-1_13

[20] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, Nicolas Roussel, and Björn Eiderbäck. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the 2003 Conference on Human Factors in Computing Systems (CHI '03)*. ACM, 17–24. https://doi.org/10.1145/642611.642616

[21] Jonas Löwgren. 2007. Interaction Design Considered as a Craft. In *HCI Remixed, Thomas Erickson and David W McDonald (eds.).* The MIT Press, Cambridge, MA. https://doi.org/10.7551/mitpress/7455.003.0041

[22] Christopher J MacLellan, Erik Harpstead, Rony Patel, and Kenneth R Koedinger. 2016. The Apprentice Learner Architecture: Closing the loop between learning theory and educational data. In *Proceedings of the 9th International Conference on Educational Data Mining (EDM '16)*. IEDMS, *151–158*.

[23] Roberto Martinez-Maldonado, Abelardo Pardo, Negin Mirriahi, Kalina Yacef, Judy Kay, and Andrew Clayphan. (2015). LATUX: An iterative workflow for designing, validating and deploying learning analytics visualisations. *Journal of Learning Analytics (JLA)*, 2(3), 9-39. https://doi.org/10.18608/jla.2015.23.3

[24] Mark W. Newman, Mark S. Ackerman, Jungwoo Kim, Atul Prakash, Zhenan Hong, Jacob Mandel, and Tao Dong. 2010. Bringing the field into the lab: supporting capture and replay of contextual data for the design of context-aware applications. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology (UIST '10)*. 105-108. https://doi.org/10.1145/1866029.1866048

[25] Amy Ogan, Evelyn Yarzebinski, Patricia Fernández and Ignacio Casas. 2015. Cognitive tutor use in Chile: Understanding classroom and lab culture. In *Proceedings of the International Conference on Artificial Intelligence in Education (AIED'15).* Springer, Cham, 318-327. https://doi.org/10.1007/978-3-319-19773-9_32

[26] William Odom, John Zimmerman, Scott Davidoff, Jodi Forlizzi, Anind K. Dey, and Min Kyung Lee. 2012. A fieldwork of the future with user enactments. In *Proceedings of the 2012 Designing Interactive Systems Conference (DIS '12)*. ACM, 338-347. https://doi.org/10.1145/2317956.2318008

[27] Luc Paquette, Ryan S. Baker, and Michal Moskal. 2018. A system-general model for the detection of gaming the system behavior in CTAT and LearnSphere. In *Proceedings of the International Conference on Artificial Intelligence in Education (AIED'18)*. Springer, Cham, 257-260. https://doi.org/10.1007/978-3-319-93846-2_47

[28] Steven Ritter, John R. Anderson, Kenneth R. Koedinger, and Albert Corbett. 2007. Cognitive Tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, 14(2), 249-255. https://doi.org/10.3758/BF03194060

[29] Susan Leigh Star. 2015. The structure of ill-structured solutions: Boundary objects and heterogeneous distributed problem solving. *Boundary objects and beyond: Working with Leigh Star*. 243-259. https://doi.org/10.1016/B978-1-55860-092-8.50006-X

[30] Ari Schlesinger, Kenton P. O'Hara, and Alex S. Taylor. 2018. Let's Talk About Race: Identity, Chatbots, and AI. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18).* ACM, 1–14. https://doi.org/10.1145/3173574.3173889

[31] Donald A Schön. 1992. Designing as reflective conversation with the materials of a design situation. *Research in Engineering Design* 3, 3: 131–147. https://doi.org/10.1007/BF01580516

[32] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 7587: 484–489. https://doi.org/10.1038/nature16961

[33] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18).* ACM, 1-14. https://doi.org/10.1145/3173574.3174014

[34] David A Wroblewski. 1991. The Construction of Human-Computer Interfaces Considered as a Craft. In *Taking Software Design Seriously.* Academic Press, 1–19.

[35] Qian Yang, Justin Cranshaw, Saleema Amershi, Shamsi T. Iqbal, and Jaime Teevan. 2019. Sketching NLP: A Case Study of Exploring the Right Things to Design with Language Intelligence." In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*, ACM, 1-12. https://doi.org/10.1145/3290605.3300415

[36] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. To appear in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*.

[37] John Zimmerman and Jodi Forlizzi. 2017. Speed dating: providing a menu of possible futures. *She Ji: The Journal of Design, Economics, and Innovation*, 3(1), 30-50. https://doi.org/10.1016/j.sheji.2017.08.003

[38] John Zimmerman and Jodi Forlizzi. 2019. Service design. In *The Encyclopedia of Human-Computer Interaction 2nd Edition* (53). Interaction Design Foundation.