Dancing with the AI Devil: Investigating the Partnership Between Lawyers and AI

Mary Mikhail, Adam Roegiest, Karen Anello
firstname.lastname@kirasystems.com
Kira Systems
Toronto, Canada

Winter Wei
University of Toronto
Toronto, Canada
winterwei@gmail.com

ABSTRACT

As professional users interact with more AI-enabled tools, it has become increasingly important to understand how their work and behaviour are affected by such tools. In this paper, we present the insights that we have gleaned from a qualitative user study conducted with nine of our software's users who are all legal professionals. We find that as our participants become more accustomed to the system they begin to subtly alter their behaviours and interactions with the system. Using their shared experiences, we distill these into insights that may inform the design of similar systems.

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**; • **Computing methodologies → Machine learning**;

ACM Reference Format:
Mary Mikhail, Adam Roegiest, Karen Anello and Winter Wei. 2020. Dancing with the AI Devil: Investigating the Partnership Between Lawyers and AI. In *2020 Conference on Human Information Interaction and Retrieval (CHIIR '20), March 14–18, 2020, Vancouver, BC, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3343413.3378014

1 INTRODUCTION

> *"When you dance with the devil, the devil doesn't change. The devil changes you."*
>
> Amanda Hocking

There is an increasing trend to investigate the ability of users to explain and understand AI-powered systems [1, 2, 5, 21, 22]. Explainable AI [8] has inspired a user-centric perspective in designing intelligible AI systems [1, 18]. While some offer guidelines [2] into how systems ought to be designed to facilitate explainability, understandability, and usability of such systems; others [10, 15, 24] seek to explore these topics from the users perspective; and others promote the investigation into coupling behavioural research with explainable systems [1, 18]. Motivated by these research trajectories and an existing body of research in the legal domain [3, 11, 12, 16], we present an investigation into how legal professionals understand and adapt their behaviour when using our AI-powered system to train and evaluate models to extract information from documents.

We seek to understand how our users develop and build their sense of proprioception,[1] which refers to our sense of body position and self-movement, in relation to our AI-powered system. To help ground our story we use a motivating example of two dancers and their interactions as both must have developed this sense to successfully dance. Through semi-structured interviews with nine participants drawn from our user population, we seek to understand how they learn to train our system, refine its effectiveness, and how their behaviour is changed over time.

To begin, we first describe the underlying system used by our participants in their day-to-day workflow to provide sufficient context into the functionality of the system. We then provide an overview of our study methodology, demographics, and coding practice for the semi-structured interviews. Having coded our interviews, we detail the various high-level insights that we gleaned from our coding process and use our dance metaphor to help ground the insights. In particular, we focus on how participants learn to train the system, use the system's false positives and negatives to guide refinements, and use their understanding of the system to begin refining their document annotation strategies to accommodate for the behaviour of the system. From these observations, we posit implications for the design of systems that partner human and AI.

2 OUR SYSTEM

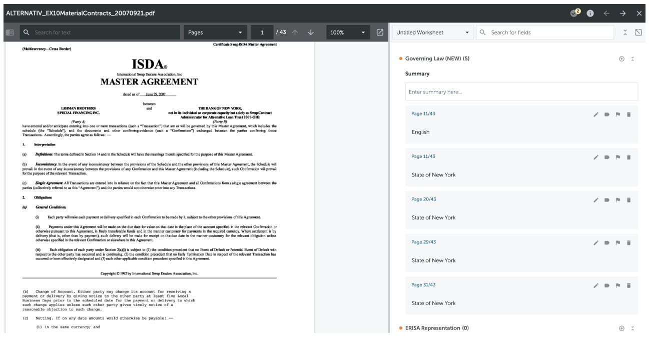
Our system, Kira, whose precise details have been described elsewhere [6, 7], allows users to upload and review documents, extract desired topics and concepts (e.g., dates, parties, provisions), and annotate documents for new concepts to facilitate training new machine learning models to extract that content. The primary goal of the system is to allow users to efficiently review documents for the purposes of legal due diligence [17, 19], audits, and other matters. In Figure 1, we show a snapshot of the main interface with which users interact to perform the review, extraction, and annotation on those documents. We have found that our users have varying workflows, with some reviewing and extracting in order to generate a final report or deliverable; in contract, others prefer to continually add on to their Kira instance (e.g., for use as a clause library).

The crux of our solution is not in the document viewer but in the ability to self-train machine learning models to identify desired topics and concepts. While this functionality was initially an internal proof of concept, when provided to users it quickly became a "go to" feature. Accordingly, it has been developed to allow users to "capture their expertise in the system." This feature very quickly

CHIIR '20, March 14–18, 2020, Vancouver, BC, Canada

ACM ISBN 978-1-4503-6892-6/20/03…$15.00
https://doi.org/10.1145/3343413.3378014

[1] Our use of proprioception in the context of AI systems is inspired by Jeremy Pickens.



Certificate Swap ISDA Master Agreement

(Multicurrency—Cross Border)

ISDA®

International Swap Dealers Association, Inc

MASTER AGREEMENT

dated as of June 29, 2007

between

LEHMAN BROTHERS SPECIAL FINANCING INC.
(Party A)

and

THE BANK OF NEW YORK, not in its individual or corporate capacity but solely as Swap Contract Administrator for Alternative Loan Trust 2007-OH2
(Party B)

have entered and/or anticipate entering into one or more transactions (each a "Transaction") that are or will be governed by this Master Agreement, which includes the schedule (the "Schedule"), and the documents and other confirming evidence (each a "Confirmation") exchanged between the parties confirming those Transactions. Accordingly, the parties agree as follows: —

1. Interpretation

(a) *Definitions.* The terms defined in Section 14 and in the Schedule will have the meanings therein specified for the purpose of this Master Agreement.

(b) *Inconsistency.* In the event of any inconsistency between the provisions of the Schedule and the other provisions of this Master Agreement, the Schedule will prevail. In the event of any inconsistency between the provisions of any Confirmation and this Master Agreement (including the Schedule), such Confirmation will prevail for the purpose of the relevant Transaction.

(c) *Single Agreement.* All Transactions are entered into in reliance on the fact that this Master Agreement and all Confirmations form a single agreement between the parties (collectively referred to as this "Agreement"), and the parties would not otherwise enter into any Transactions.

2. Obligations

(a) *General Conditions.*

(i) Each party will make each payment or delivery specified in each Confirmation to be made by it, subject to the other provisions of this Agreement.

(ii) Payments under this Agreement will be made on the due date for value on that date in the place of the account specified in the relevant Confirmation or otherwise pursuant to this Agreement, in freely transferable funds and in the manner customary for payments in the required currency. Where settlement is by delivery (that is, other than by payment), such delivery will be made for receipt on the due date in the manner customary for the relevant obligation unless otherwise specified in the relevant Confirmation or elsewhere in this Agreement.

(iii) Each obligation of each party under Section 2(a)(i) is subject to (1) the condition precedent that no Event of Default or Potential Event of Default with respect to the other party has occurred and is continuing, (2) the condition precedent that no Early Termination Date in respect of the relevant Transaction has occurred or been effectively designated and (3) each other applicable condition precedent specified in this Agreement.

Copyright © 1992 by International Swap Dealers Association, Inc.

```
(b)  Change of Account. Either party may change its account for receiving a
payment or delivery by giving notice to the other party at least five Local
Business Days prior to the scheduled date for the payment or delivery to which
such change applies unless such other party gives timely notice of a
reasonable objection to such change.

(c)  Netting. If on any date amounts would otherwise be payable: --

     (i) in the same currency; and
```

Governing Law (NEW) (5)

Summary

Enter summary here...

Page 11/43
English

Page 11/43
State of New York

Page 20/43
State of New York

Page 29/43
State of New York

Page 31/43
State of New York

ERISA Representation (0)

Figure 1: A screenshot of our document viewer that is used to review identified content, highlight relevant content, and help train the system to identify new concepts.

allows experts to train custom models to identify concepts that consistently surprise us (e.g., identifying win conditions in board game manuals). A user's workflow influences the amount and type of risk they are willing to take when reviewing documents. For users with projects that require the most rigorous human review, they may deliberately use Quick Study to produce under-inclusive models to avoid the distraction of false positives. However, other users may use Quick Study to provide over-inclusive models to ensure that all relevant information is identified and can be correctly assessed. Accordingly, these different uses of our system motivates our desire to understand how our non-technical users understand and use this feature to train their own high-quality models.

In Figure 2, we present the two primary tabs of the Quick Study feature. The default tab is the "learning" tab (Figure 2a) which allows users to start model training (not depicted) and provides summary data on model effectiveness[2]. These summary measures were selected with the intent of being easy to interpret "true hits" (i.e., true positives), "false hits" (i.e., false positives), "misses" (i.e., false negatives),Recall, Precision, and the F1 score. In isolation, these measures do not allow users to take corrective action and improve model effectiveness and so, we provide a "validation" tab (Figure 2b). The tab allows users to view (dis)agreements with the system. While the extractions are presented in isolation, they can be clicked through to view the extraction in context. These tabs allow users to train, assess, and refine models to achieve their desired accuracy.

3 STUDY METHODOLOGY

Prior to our study we conducted a round of pilot research with three of our in-house lawyers that extensively use Kira and Quick Study to produce models that are disseminated to our clients. This pilot was used to help focus our subsequent research sessions with external users of the system. Our main study consisted of two rounds of semi-structured interviews that probed deeply into how users viewed their training of the system. During our interviews, we sought insight into participants' workflows and processes they have in place to effectively use Kira and Quick Study. We also asked participants to summarize their understanding of the information presented in the learning and validation tabs and how they use this information. To aid participants in formulating concrete thoughts, we guided their discussion using a previously trained model that initially exhibited unexpected results or behaviour. All sessions used the same interview guide, where participants were asked two Likert scale questions and the rest were open-ended. Sessions were conducted remotely, recorded, and transcribed for analysis.

For our study we recruited nine participants (2 women, 7 men) from our client user base. On average, these participants had two years of experience with the system as a whole. Five participants have a legal background and the remainder have domain expertise in other fields. Their backgrounds are representative of our users demographics. There was no special selection criterion for participation except that they use the system to train models.

Using the transcriptions, the first author used thematic analysis [9, 23] to identify and provide a coding schematic that was used to analyze participant comments. Affinity diagrams were used to cluster information generated in the coding phase to glean additional insights. Our open-coding revealed themes including the refinement of annotation strategies based on outputs of the model, the understanding of how documents impact validation, and how variations in documents can affect trained models. In the following sections, we explore these themes more fully.

4 RESULTS AND DISCUSSION

In this section, we break down the phases that our participants expressed going through as they gained familiarity with our system.

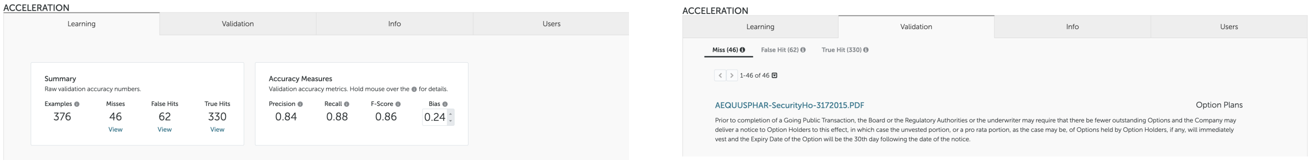
4.1 Learning to Train

When learning a new dance, whether its the Texas Two-Step, the Lindy Hop or Salsa, teachers will often try to build on the existing understanding that a learner may have (e.g., step counts, movements). In so doing, they seek to build upon a shared understanding. We find that similar behaviour is exhibited by our participants. Domain expert participants (e.g., lawyers) will often use existing educational material to teach themselves a new concept (e.g., new GDPR legislation). Non-experts, on the other hand, will instead seek advice and understanding from the attorneys that they work alongside. While the mechanism of learning may be different, all participants build upon their existing knowledge to gain insight into how to identify new relevant material in training documents.

As they continue to refine and improve their models, our participants reported the need to understand more nuanced aspects of the system in order to create cohesion between their needs and the limitations of the system. As one user suggests, *"Because users will say 'I want to extract every time it says the word 3 percent deduction' and that's not what they want to do. That's not true right. You have to distill that into its essential oil which is "I need to know every time this provision is changed in this way'."* Indeed, this hearkens back to Blair and Moran's seminal study [3] on legal keyword search, which found that paralegals would continually refine queries with lawyers to more accurately capture the information need.

To successfully train a model, participants conduct iterative processes to improve on their initial understanding of a concept because they find *"you need consistency to be able to train a model."* Eight of our participants shared the idea that some concepts (e.g., legal definitions) are as easy for the system to identify as they are a human. This is facilitated by agreement between trainer and system

[2] As assessed through the 5-fold cross validation process described in [7].



Summary

Raw validation accuracy numbers.

Examples	Misses	False Hits	True Hits
376	46	62	330
	View	View	View

Accuracy Measures

Validation accuracy metrics. Hold mouse over the ⓘ for details.

Precision	Recall	F-Score	Bias
0.84	0.88	0.86	0.24

(a) The learning tab which depicts summary measures of the underlying model's effectiveness.

AEQUUSPHAR-SecurityHo-3172015.PDF Option Plans

Prior to completion of a Going Public Transaction, the Board or the Regulatory Authorities or the underwriter may require that there be fewer outstanding Options and the Company may deliver a notice to Option Holders to this effect, in which case the unvested portion, or a pro rata portion, as the case may be, of Options held by Option Holders, if any, will immediately vest and the Expiry Date of the Option will be the 30th day following the date of the notice.

(b) The validation tab allows users to view the true positives, false positives, and false negatives of the model.

Figure 2: Examples of the machine learning tabs that allow users to review how well their models have been trained.

on strict formatting and acceptance criteria (e.g., being encapsulated in parentheses). However, as a concept gains added complexity some begin to ask *"what are all the things I would need as a human to make a decision with this extraction?"* and in so doing start to break the concept down into *"narrower bits of things."* This reinforces the results of Fiebrink et al. [5] which suggest that the discovery and management of the trade-offs in building models that are easy to train and understand is a dynamic and iterative process.

When starting with a new partner, a dancer may attempt to anticipate actions and reactions based upon some preconceived understanding. Our participants consistently expressed a similar tendency when attempting to teach the system a new concept. A participant expressed that they want to provide examples *"saying that you want to find a paragraph like this, not necessarily exactly like this but a paragraph like this."* To accommodate this they often annotate more text in examples than they would for a human so that they capture more context. Much like a new dancer, this may work but relies heavily on their partner to accommodate their errors. Similarly, participants learn over time when providing more context is helpful and when it might lead to over inclusivity.

As a users experience grows, we believe that they are able to build a sense of proprioception in conjunction with Kira. This allows them to understand how differences in layout, tables, lists, and OCR errors can all effect the outcome of the model. Lewis and Williams called this process "experiential education" [14] whereby immersion into the experience and subsequent reflection allows learners to continually develop. Our participants use a similar process to help them identify what changes can be made to improve their models. Over time it seems that users learn how the system can respond to different inputs and how to correct undesirable behaviour.

4.2 Changing Roles

As a dancer develops, they may find that they switch from following to leading the routine or vice versa. Our participants expressed that as they began to trust and understand how the system operates at a high-level, they began to allow the system to lead the corrective actions being taken. By using the summary information (Figure 2a) and concrete examples (Figure 2b), users are guided to aspects of the concept over which the system might struggle (i.e., false positives/negatives) and those that it "understands" (i.e., true positives). Participants can then take action to add, remove, or refine training annotations to guide Kira in a given way.

Perhaps unsurprisingly, participants expressed difficulty in understanding the exact meaning behind accuracy scores (e.g., Precision) but understand that as a measurement of agreement between the system's understanding of a concept and their own. In interviews, participants made it clear that these scores were not seen as an *"absolute"* but were reflective of what participants taught the system and the materials with which it was taught (e.g., it may not work well on unseen document types). Participants' shared the view that scores were reported in the context of *"Kira's thinking"* and that once data is trained on, it is *"Kira's knowledge."* In this way, participants highlighted distinctions between their teachings, and how Kira learns and understands the concept in question.

Despite struggling at times to grasp the meaning of accuracy scores, our participants unanimously perceive comfort in the numbers themselves as they are able to provide an indication of what the system *"might be missing or what it's struggling with."* These scores are used by participants as a means to measure performance, reliability, and stability. Indeed, some compare it to *"a student with that grade taking a test today"* and finding comfort in the associated (test) score. Accordingly, they rely on *"spot checks [to] make sure you're not overly reliant on Kira"*. A user's corrective actions then helps to further align the user to the system and vice versa which leads to increased proprioception.

4.3 Refining the Routine

Long term dance partners will develop and refine their routine to leverage better techniques and improve flow. While our participants suggested that the scores help guide their refinement, they found that examination of the false positives and negatives was very critical to analyzing the granular differences in system's understanding and their own. As one participant described it, *"although it doesn't direct you...it's help enough to see what Kira thinks is common."* In particular for false positives, all participants describe scanning the extractions and their ability to identify subtle differences as means of expediting model validation.

By examining differences, participants reported that their highlighting strategy changes over time. Seeing validation details allows them to explore ways in which they can refine and improve their highlighting strategy given how they view the system. This may reflect a symbiotic relationship in the learning process whereby our participants are changed by Kira as much as as they change Kira. As one participant explains, "*[by] highlighting the ones where the dollar amount's in the title it still captures the titles where there's no dollar amount in there. And so you ask yourself a question: Do I have to add this in here?"* This continual refinement of the training data and how participants view the system and its ability to understand concepts may yield improvements in models trained in the future.

We observed a general consensus that participants often attempted to look for different semantic patterns as to why some annotations were false negatives and others were correctly identified. False negatives that appear at a high-level to be similar to true positives are seen as motivators for strategic readjustments of their training strategy. One participant recounts that to deal with the differences caused by OCR variances, they had to *"[write] down the ones that would work and the ones that didn't."* By doing this, they were able to identify documents exhibiting certain problematic properties that they could then add to the training data to improve the effectiveness of the model.

As participants grew accustomed with Kira, they began to anthropomorphize Kira and their interactions. One participant expressed the idea that it's *"how Kira thinks I shouldn't have highlighted...So it's telling me how good I was at highlighting and not missing things that Kira thinks I should have highlighted...there are probably only a few items that Kira thinks I should have highlighted that I didn't."* It would seem that that participants seek to help the system support them and determine how they can help support the system.

4.4 Chasing Perfection

While all dancers may seek to excel at particular style or routine, some will aspire to attain perfection and will work hard to achieve it. For those of our non-power user participants that only train models when necessary, they eventually reach a point where models become fit for purpose. Subsequent refinement is no longer needed as any deficiencies can be mitigated elsewhere in their workflow. On the other hand, a portion of power users want to eek as much potential out of the system as possible.

Power users will leverage their ability to perturb the model's bias, which merely adjusts how permissive or restrictive the model is when highlighting content. By increasing the bias, users can force the model to be more permissive and can examine what is on or near the decision threshold (how "far away" false negative are) and this works similarly for false positives when decreasing the bias. Regardless of the exact bias value, adjusting the bias allows them to find new patterns that can help inform what new information they may use to train the model.

5 LOOKING FORWARD

Our participants seek to translate their conception of relevance into a form the system can understand and then seek to optimize this translation process to achieve their ultimate goals. How we facilitate this translation is on us as the system designers and builders. The mechanisms in which we display feedback to users and how we allow them to act on that feedback is critical to facilitating useful interactions. Too little and users may just attempt a "guess and check" approach but too much may cause information overload. Essentially, a user must be able to make sense of what the system is learning and formulate falsifiable hypotheses as to why. While the user may not ever (desire to) understand the underlying mechanics of the system, they should be able to generate some form of proprioception around their interactions with the system; much as dancers do as they improve at their craft.

Cohen and Feigenbaum[4] suggest that "the trick is to know enough about how humans and computers think to say exactly what they have in common, and, when we lack this knowledge, to use the comparison to suggest theories of human thinking or computer thinking." From our participants, we've seen nuanced interaction in how our users interact with the system and how they perceive the system's "thinking." Moreover, the interactions between user and system change the resulting behaviours of both. Whether this is in the model produced at the end of the process or the manner in which a user annotates documents. By further attempting to understand what users believe to be true of a system (e.g., how their actions produce different effects), we may arrive at an understanding of how better to design interactions that are useful rather than those that are harmful.

Guidline 16 of by Amershi et al.'s[2] Human-AI interaction guidelines, states that systems should "[c]onvey the consequences of user actions. Immediately update or convey how user actions will impact future behaviors of the AI system." While our participants use pattern matching and their intuitive understanding of Quick Study to inform corrections, this is not an immediate update nor does it always provide actionable knowledge. While we could preemptively produce the model with the correction applied and display the changes to the model, the computational cost may render this infeasible at a large-scale. Accordingly, we need to consider alternatives that illustrate potential changes to the system. Whether this is creating estimates around effectiveness (e.g., "changing these examples *may* yield this much change in Recall") or to help guide corrections to be most impactful (e.g., through uncertainty sampling [13] to improve the decision boundary).

While our participants and many of Kira's users are experts in their fields, they are not experts in ML evaluation. Accordingly, they create narratives to help their understanding of what evaluation measures mean to them. As we discussed in Section 4.2, our participants often frame these measures in terms of agreement between themselves and the system. For example, we may recast Recall as "the proportion of examples that the system agreed with the user," which while not the most technically accurate may help to better influence their understanding of how the system is behaving. By building intuition of how the system behaves in response to their actions (i.e., developing proprioception), we may foster more positive interactions while minimizing negative ones. This goal is reinforced by customer experience research [20] that has shown positive interactions can make users willing to forgive mistakes.

6 CONCLUSION

Throughout this work we highlight the interplay between system and user and how there are transformations between both parties. Yet, the system does not change as a matter of course, the resulting model does. Accordingly, our system takes the role of the "devil" in Hocking's quote from Section 1. Our participants state how they change their behaviour in response to the system but only what the system returns to the user is changed. The underlying mathematics of the system do not change. In this way, we must tread with great care in how we build systems that affect user's lives and livelihoods. By creating systems that have the potential to change how users behave, even when this is ostensibly benign, may have unintended consequences (e.g., chasing higher accuracy scores rather than a robust model). Outside of making our users more efficient or effective, the systems we build may very well change how our users behave and with that power comes a responsibility to wield it with care and understanding.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proc. CHI 2018*. ACM.

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction *(CHI '19)*.

[3] David C. Blair and M. E. Maron. 1985. An Evaluation of Retrieval Effectiveness for a Full-text Document-retrieval System. *Commun. ACM* (March 1985).

[4] P.R. Cohen and E.A. Feigenbaum. 2014. *The Handbook of Artificial Intelligence*. Number v. 3. Elsevier Science.

[5] Rebecca Fiebrink, Perry R. Cook, and Dan Trueman. 2011. Human Model Evaluation in Interactive Supervised Learning *(CHI '11)*.

[6] Blinded for Review. [n. d.].

[7] Blinded for Review. [n. d.].

[8] David Gunning. 2017. Explainable Artificial Intelligence (xai). *Defense Advanced Research Projects Agency (DARPA)* 2 (2017).

[9] Erika Hall. 2013. *Just Enough Research*. A Book Apart.

[10] Dagmar Kern, Daniel Hienert, Katrin Angerbauer, Tilman Dingler, and Pia Borlund. 2019. Lessons Learned from Users Reading Highlighted Abstracts in a Digital Library. In *Proc. CHIIR 2019*. ACM.

[11] Ben Klaber. 2013. Artificial Intelligence and Transactional Law: Automated M&A Due Diligence. In *ICAIL DESI V Workshop*.

[12] Gloria J Leckie, Karen E Pettigrew, and Christian Sylvain. 1996. Modeling the information seeking of professionals: A general model derived from research on engineers, health care professionals, and lawyers. *The Library Quarterly* 66, 2 (1996).

[13] David D. Lewis and William A. Gale. 1994. A Sequential Algorithm for Training Text Classifiers *(SIGIR '94)*.

[14] L. Lewis and C. Williams. [n. d.]. Experiential learning: Past and present. New Directions for Adult & Continuing Education. *1994(62), 5-16.* ([n. d.]).

[15] Brian Y Lim and Anind K Dey. 2011. Design of an intelligible mobile context-aware application. In *Proc. MobileHCI 2011*. ACM.

[16] Stephann Makri, Ann Blandford, and Anna L Cox. 2008. Investigating the information-seeking behaviour of academic lawyers: From Ellis's model to design. *Information Processing & Management* 44, 2 (2008).

[17] Jeffrey Manns and Robert Anderson. 2017. Engineering Greater Efficiency in Mergers and Acquisitions. 72 (2017).

[18] Tim Miller. 2018. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* (2018).

[19] James A Sherer, Taylor M Hoffman, and Eugenio E Ortiz. 2015. Merger and Acquisition Due Diligence: A Proposed Framework to Incorporate Data Privacy, Information Security, E-Discovery, and Information Governance into Due Diligence Practices. *Rich. JL & Tech.* 21 (2015).

[20] Bruce Temkin. 2017. Customer Experience Leads to Recommendations (Charts For 20 Industries). https://experiencematters.blog/2017/02/23/customer-experience-leads-to-recommendations-charts-for-20-industries/. (23 Feb. 2017).

[21] Joe Tullio, Anind K. Dey, Jason Chalecki, and James Fogarty. 2007. How It Works: A Field Study of Non-technical Users Interacting with an Intelligent System *(CHI '07)*.

[22] Justin Weisz, Mohit Jain, Narendra Joshi, James Johnson, and Ingrid Lange. 2019. BigBlueBot: teaching strategies for successful human-agent interactions. In *Proc. IUI '19*.

[23] Carla Willig. 2012. *Introducing Qualitative Research in Psychology*. Open University Press.

[24] Rayoung Yang and Mark W Newman. 2013. Learning from a learning thermostat: lessons for intelligent systems for the home. In *Proc. Ubicomp 2013*. ACM.