Identifying Critical Contextual Design Cues Through a Machine Learning Approach

Mary L. “Missy” Cummings
Alex Stimpson
Duke University

Introduction

The development of autonomous technologies that take on safety critical functions, such as driverless cars or surgical robots, can potentially reduce accidents and errors and improve productivity. However, while autonomous systems show promise for enhancing safety and productivity, previous research in human-automation interaction has demonstrated that adding automation does not necessarily guarantee increased system effectiveness or safety. Often, automating a task within a larger system modifies the task by transferring the operator’s workload from one physical or cognitive resource to another, thereby changing the task rather than improving it (Bainbridge, 1987). Poorly designed automation that is not understood by operators often causes human error and reduces system effectiveness due to “clumsy” implementations (Lee & Morgan, 1994).

As these systems proliferate, there is an increasing need to understand how such systems should be designed to promote effective interactions between one or more humans working with or around autonomous systems. This is especially true for safety critical settings like operators of such systems including medical systems, factory workers engaged in tasks with or near automation, or pedestrians and bicyclists operating in the same environment with driverless cars. Given the importance of promoting effective and safe interactions between human users and autonomous systems, designers of these systems need tools that allow them to determine not just which designs are effective, but how such systems fare under different contexts. Indeed, the ability of autonomous systems to account for context and changing environments is a significant hurdle that limit applications (Daily, Medasani, Behringer, & Trivedi, 2017; Marcus, 2018; Srinivasan, 2016).

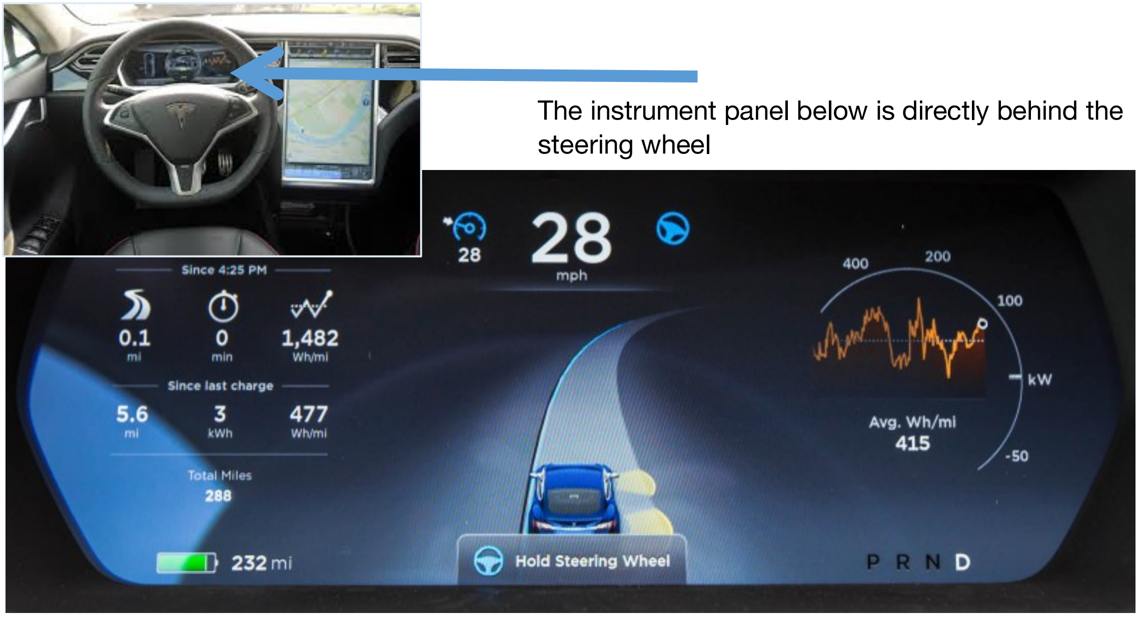
One often overlooked source of potential contextual design cues in autonomous systems is the vast amount of data generated by these systems, including those of user interactions. While machine learning approaches to data analysis are often touted for their importance in the operation of these cars, they can also be harnessed for understanding the impact of context, particularly when attempting to determine the effectiveness and safety of a design choice. This paper will discuss the importance of contextual cues in design and demonstrate how machine learning method can be adapted to determine the effectiveness of design features in an autonomous system.

Case Study: The Tesla Human Interface

One recent example of a major design flaw in terms of human-robot interaction is that of the Autopilot display in the Tesla Model S. The Tesla Autopilot is billed as a driver-assist technology, in that once engaged, the car can track itself automatically between lane lines, change lanes, and brake and accelerate as needed to move through traffic. The driver is relieved of direct control but must monitor the evolving driving situation both internal and external to the car, and ensure the car behaves in a safe manner. Figure 1 illustrates the Autopilot features. The faint blue lines on the road icon indicate that Autopilot sees the lane markings. The blue icon to the left of the speed indicates that TACC (Traffic Aware Cruise Control) is active, and holding the driver-requested speed of 28 mph. The blue steering wheel to the right of the speed indicates that Autosteer is engaged with slight right turn control input.

Despite the Autopilot’s billing as a driver-assist system, human drivers with no formal training must not only watch outside the car for possible problems, but they are supposed to cross check what they see outside with the display on the inside of the car to ensure Autopilot is working as advertised. Given that this is a complex cognitive task with no formal training, it is not surprising that there have been multiple reports of Tesla crashes where drivers did not understand that the autopilot was attempting to alert them to put their hands on the wheel (Crowe, 2016). Figure 1 illustrates what this relatively subdued alert looks like, which is the small message at the bottom of the display that says “Hold Steering Wheel”. The inset of the car cockpit shows the relative size of the instrument cluster, which is clearly not the most salient display in the car.

Such problems where human operators of complex automated systems are confused by the displays and do not understand the communications from the automation are widespread, occurring in domains such an anesthesiology (Ruskin et al., 2013), process control (Guerlain, Jamieson, Bullemer, & Blair, 2002), pilots of commercial aircraft (Vakil & Hansman, 2002), and military weapons systems (Cummings, 2004). This problem is so well documented that it has been termed "mode confusion", which occurs when an operator's mental model differs from the automation's behavior (Lankenau, 2002). Mode confusion can lead to errors when the operator responds incorrectly to the automation, thinking it is in a different state (Norman, 1983).



The instrument panel below is directly behind the steering wheel

Figure 1: Tesla Model S Instrument Display Reminding Driver to Put Hands on the Steering Wheel

Despite the clear link between how information is communicated via a display to a system user and the ability of the user to correctly interpret the system's actions, designers of such systems still struggle to develop displays that promote effective interactions with relevant stakeholders. In an ideal setting, design of such displays focuses on developing a clear mapping between user goals and the execution of those goals (Norman, 1986). A key design parameter for such goal mapping dictates that those cues in a display (visual, aural, or haptic) be designed for saliency, i.e., making sure those environmental cues that are critical to decision making are available and prominently displayed (Wiener & Nagel, 1988). Moreover, such cues should be consistent with a user's mental model (Wickens & Hollands, 2000).

Significant past research has shown that contextual cues are an important aspect of user interface design (Cockburn, Karlson, & Bederson, 2009; Lesch, Powell, Horrey, & Wogalter, 2013; Nielsen, 1993; Norman, 1988; Schneiderman, 1987), particularly in safety critical systems (Feary et al., 2013; Franke, Daniels, & McFarlane, 2002; Wei, 2014). Moreover, understanding contextual cues is critical for developing accurate system requirements (Wenxin & Kekang, 2008), since desired human behaviors cannot be achieved if the correct cues are not identified early on in the design process.

Thus, intended design cues for autonomous system operation, in theory, communicate information to the user about a system state that requires human attention and response. However, in such complex systems with multiple displays that include visual, aural and sometimes haptic (like vibrating seats), there exists inadvertent exogenous cues from the world (e.g., a cell phone vibrates in the car capturing the driver's attention) or internal stimuli (reaching a state of boredom that motivates a driver to search for an enjoyable radio station) that cause operators of complex systems to miss intended design cues, and thus potentially put themselves, the physical systems, and others at risk.

Given the importance of both identifying and designing the right contextual cues for autonomous system operation, a central design question for such systems, particularly those that are safety critical like cars and manufacturing or surgical robots, is how designers know which cues are the right ones to emphasize in a display and when? For example, how does a designer of the display in Figure 1 know that the message at the bottom of the display or an accompanying aural warning is likely not effective?

Other critical related design questions include how exogenous cues influence the perception of intended design cues, such as road signage and other displays in the vehicle like the large map display in the inset of Figure 1. And while the investigation of cue interpretation of individual behaviors provides useful insight, designers of mass consumer products need to understand potentially large population effects such as the role of culture, age and experience. Thus, there is a need for a design evaluation method that can analyze large data sets and identify individual interaction strategies, but also the influence of potentially secondary variables like demographic characteristics.

Currently, designers of interfaces for autonomous systems evaluate how well their cue selections match mental models and align with system needs by conducting surveys, focus groups, individual usability testing, and on occasion, statistical hypothesis testing which typically takes the form of A vs. B testing, i.e. testing two competing versions of a design to determine which display more often produces the desired behavior (Nielsen, 2005). While such methods contribute to a designer's understanding, other than inferential testing, most of these methods are highly subjective. While useful for understanding preferences, subjective evaluations may not address the true effectiveness of intended design cues.

It has long been established that people are generally not effective at determining what cues influence their judgments (Kraut & Lewis, 1982) (Andre & Wickens, 1995; Nisbett & Wilson, 1977; Wilson & Nisbett, 1978). As a result, designers who elect to use focus groups and subjective surveys to assess their designs for cue salience are likely to obtain inaccurate results. Moreover, while hypothesis-driven tests like A vs. B testing provide objective results, they are often costly to develop, and obtaining statistical significance can be difficult without large sample sizes. Moreover, the hypotheses are typically very narrow by design (Nielsen, 2005), and the ability to see the interaction of various factor effects is often lost as part of the focus on minimizing model error to increase model fit.

Thus, what is needed is an analytical strategy that designers of autonomous systems can use to determine whether their designs are indeed capturing their planned human-system design cues, as well as the impact of potential interference from exogenous cues or internal stimuli. Moreover, since autonomous systems often generate significant amounts of data, a useful design analytic tool is one that allows industry engineers the ability to leverage significant amounts of data at their disposal, as well as allows them to more meaningfully analyze small data sets. Such an analytical approach should be able to capture both individualized outcomes as well as more global effects due to a design factor under consideration. Most importantly, such an approach should be usable and understandable by industry designers who are both time and budget pressured.

Design Applications of Machine Learning

A popular data analytic methodology commonly used in the design of autonomous system software is machine learning, which is a methodology that attempts to automate analytical model building through automatic discovery of regularities in data that can be used to classify the data into different categories (Bishop, 2006). While widely used in a number of fields like computer vision and voice recognition systems, in terms of designing for human interaction, machine learning is a relatively new approach. Recent advancements in generative design apply machine learning techniques to discover new designs through optimization of physical parameters (Machwe & Parmee, 1997; Yu, Pan, Matsunawa, & Zeng, 2015). However, no one has investigated how to extend such methods to the design and optimization of displays meant to promote interaction between a complex cyberphysical system and a human user.

There has been significant work attempting to model users of technology through machine learning methods (see Webb, Pazzani, & Billsus, 2001) for a review), but with little to no attention paid to how such models can inform system design. For example, Huang, Oviatt, and Lunsford (2006) developed a user model based on Machine Learning (ML) to better predict users' multimodal integration patterns via speech and a pen, in order to develop a system that dynamically responded to user behaviors. However, while they were successful in developing a predictive model, this model was never actually applied to the design of a system.

Indeed, much of the user modeling research based on ML has focused on choice preference modeling, which is widely used in marketing design decisions for websites worldwide. For example, in marketing applications, machine learning is used to design a more personalized customer or user experience through an interface, such as Amazon providing recommendations for other products a user may be interested in (Hebron, 2016). However, as Webb (2001) points out, the bulk of this research has focused on the modeling of individuals' choices/preferences and that applications of machine learning for discovering users' characteristics, which are critical design considerations, are rare.

This is an important distinction since the safety of autonomous system operators often depends on their understanding of the cues provided by the system, so understanding user preferences in safety critical systems is less important than understanding the users' use of cues, both planned and unplanned. Moreover, because practicing engineers are interested in designing for populations of users as well as individual users, they need to understand how system designs can and should be tailored for different classes of users (e.g., experts vs. novices, older vs. younger users, etc.)

This problem highlights another critical missing element in the past applications of machine learning to user modeling, which is a lack of linking user choices and behaviors to the performance of an overall system. Instead of predicting what products a person is likely to prefer, in systems that require human-autonomous system interaction, a system must be able to dynamically adapt its displays through predictions of not just the current and likely future state of the system, but also of the human. For example, self-driving cars will need to dynamically determine when to display information for timely passing of control between a self-driving car and the driver or when to signal to a surgeon of a robotic system that a dangerous limit is being approached.

Thus, models of user preferences are not enough to inform the design of such displays, but rather models are needed that consider the current and predicted state of the system and the world, as well as the ability of users to perform a task under dynamic conditions. Past research on machine learning and user performance modeling is limited, with some researchers focusing on developing machine learning models of student performance (Amershi & Conati, 2007; Stimpson & Cummings, 2014; Webb et al., 2001), driving performance (Pentland & Liu, 1999), and supervision of unmanned aerial vehicles (Boussemart, Cummings, Las Fargeas, & Roy, 2011). While these models predicted users' performance with varying degrees of success, such models were not explicitly used to consider different design options, particularly as they relate to the performance of systems with high degrees of freedom.

One unique aspect of autonomous systems that makes machine learning a potentially valuable design tool is that for most of these systems, significant data is gathered through the large number of data points generated by various sensors. For example, a self-driving car generates about 1 gigabyte of data per sec from the RADAR, LIDAR, and camera systems (Gross, 2013), including when a driver's hands are on the wheel and the magnitude of his or her control inputs, as well as when the user touches every knob and lever inside the car. Given the enormity of such data, it is very difficult to apply traditional hypothesis-driven statistical methods, especially after collection.

Such traditional hypothesis-driven models, often expressed through Analyses of Variance or other regression derivations, inherently assume a model, as opposed to algorithmic modeling, i.e., machine learning, which can be used both on large complex data sets and can be a more accurate and informative alternative to data modeling on smaller data sets (Breiman, 2001). Taking a data-driven machine learning approach to the analysis of such data can determine not just user models, but also how such models can be explicitly linked to design decisions, which could yield results not identifiable through traditional user testing methodologies.

The Need for Explainable Machine Learning Algorithms for Design

Whether ML approaches can be developed to aid in design decisions for systems requiring significant human-autonomous systems interaction is inextricably linked to how the results of such machine learning approaches are communicated to the designer. This idea of the need for "explainable artificial intelligence" is gaining in popularity as the applications of artificial intelligence (AI) in various systems have grown (DARPA, 2016). Given that machine learning techniques are deeply rooted in probabilistic reasoning, but knowing that humans struggle with understanding probabilistic models, even for experts (Tversky & Kahneman, 1974), it is important to address just how to effectively communicate the results of a ML-based analysis. This idea of explainable ML results is even more critical if we want to develop a method for engineers in the workforce to adapt in the design of autonomous systems, as they likely will have a widely varying understanding of how ML algorithms produce results.

The concept of "explainability" is rooted in explanation-based reasoning and decision making. In explanation-based reasoning, evidence is assembled into explanatory structures representing possible classifications of the evidence (Pennington & Hastie, 1993). Given impoverished data, humans use inferential strategies to piece together a holistic picture in order to make a decision. This psychological construct has several implications in terms of human collaboration with ML algorithms. One is that the explainability must be sufficient to create a relatively complete picture for the designer, as any explanatory gaps could (potentially incorrectly) be automatically filled by the human. Another is that ML algorithms could be developed to match the explanation-based approaches used by humans, so as to be more understandable. Such approaches have been explored in the past (Hair & Pickslay, 1993), but modern ML algorithms have not incorporated these strategies.

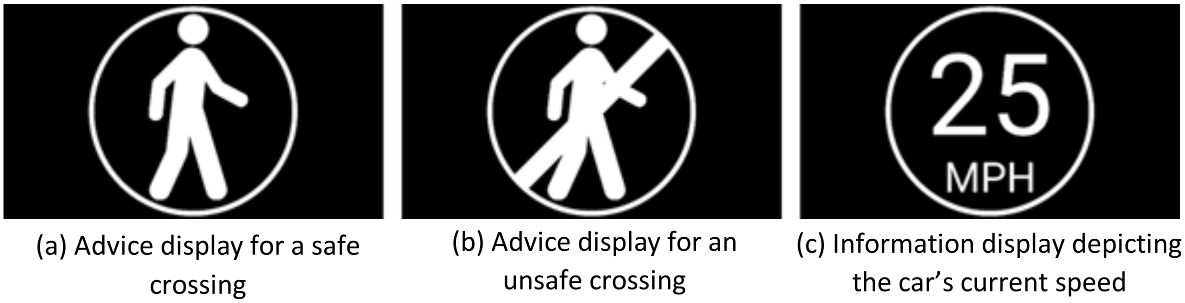
Even if it were well understood what the appropriate data was to describe the inner workings of ML algorithms, there are important considerations for what elements and what format results should be presented to a designer to support various design decisions. For example, the context of the task environment is key. ML algorithms are largely contextual (Johnson, 2014) and must be tailored to individual domains, as there is no one size fits all especially when attempting to model human behavior. In addition, the needs of designers leveraging ML algorithms will vary, as will their levels of expertise with both applying ML algorithms and their interpretation of results.

There are many potential issues with ML algorithms that make their results difficult to interpret (Ribeiro, Singh, & Guestrin, 2016). One such case is where the data used for training is not representative of the test data, i.e., data set shift (Quionero-Candela, Sugiyama, Schwaighofer, & Lawrence, 2009). Other example cases include prior probability shift, sample selection bias, and imbalanced data. Additionally, poor initialization of model structures or weights (e.g., Artificial Neural Networks, or ANNs) (Denoeux & Lengelle, 1993; Schmidt et al., 1993) or overfitting due to a lack of regularization (Moody, 1991) can result in poorly performing models. Failure to meet assumptions of algorithms such as i.i.d. (independent and identically distributed) data for Support Vector Machines (SVMs) can also result in inappropriate models (Hsu, Chang, & Lin, 2003). The "brittleness" of many ML algorithms due to sensitivities of not meeting the underlying assumptions surrounding the context and collection of data underscores the importance of explanation to a human decision maker, especially in attempting to make design decisions.

Some of these errors may be observable through prediction performance when applied to validation or test datasets, but the identification of inappropriate models and predictions can be difficult, especially without significant experience in ML. Moreover, design engineers are often time-pressured and do not have the luxury of in-depth and time-intensive sensitivity analyses, and so often will look for a "good enough" solution. Without transparency in the ML algorithm's rationale in providing a prediction, it is a challenging task for a human to understand a prediction and then translate that into a clear design choice.

Case Study: Designing Displays for Driverless Car – Pedestrian Interaction

To determine if a machine learning approach could provide additional information about contextual cues in a design setting, results from a previous study were reanalyzed using a machine learning approach. Three different displays (Figure 2) were mounted on the front grill of a simulated self-driving car to be visible to pedestrians passing in front of the car. They and a control condition of no display were tested in a human-in-the-loop experiment with 55 participants that attempted to determine whether the displays were effective in communicating a self-driving car's intent (Clamann, Aubert, & Cummings, 2017). Traditional hypothesis-driven statistical analysis led to the conclusion that the displays had no effect on pedestrians' timeliness (thus safety) of crossing decisions. The only demographic variables influential in these outcomes were participants' ages and conscientiousness scores on the NEO-FFI-3 assessment, which rates an individual's personality traits of **N**euroticism, **E**xtraversion, **O**penness to experience, **A**greeableness, and **C**onscientiousness. Older participants tended to make more safe crossing decisions than younger participants, and those who were more conscientious tended to make slower crossing decisions.



(a) Advice display for a safe crossing

(b) Advice display for an unsafe crossing

(c) Information display depicting the car's current speed

Figure 2: Proposed displays to aid pedestrians in making road crossing decisions in the presence of self-driving cars.

The results from this experiment should be interpreted in light of typical confounds such that the testing was done on a university campus with a relatively homogeneous sample with limited cultural diversity, so it is difficult to generalize these results with confidence. Moreover, the conduct of such naturalistic studies is very difficult, expensive, and time-consuming and car companies are not as interested in the design and conduct of scientifically valid studies. They simply want to know which design choices can lead to best outcomes, both from an objective and subjective perspective.

Many companies including Google, Nissan, and drive.ai have said they are going to install displays very similar to those in Figure 2 on their self-driving cars. And whether the displays are critical to pedestrian safety or instill greater trust so people feel better about sharing space with them remains to be studied. Curiously, in interviews after the pedestrian experimental trials, only 12% of participants admitted to using the displays in Figure 2, but 46% said they thought such displays should be included in the design of driverless cars. So, while there may not be clear objective performance data supporting the use of such displays, subjectively they may provide value and it is still an unknown as to how long-term exposure to these technologies could eventually influence pedestrians' decisions.

So how could the data from this experiment be analyzed to provide more useful insight concerning contextual cues? Each person conducted 12-16 trials, which resulted in a data set of 850 observations, so for this kind of experiment, there was a relatively high number of data points. As a result of this data set characteristic, we attempted to apply machine learning algorithms to determine if there were any other useful relationships that could be derived from the data. It was in doing this that we realized that much more work is needed in determining how to apply machine learning algorithms to design problems.

Difficulties in choosing the best algorithm

In order to be useful in a design context, machine learning approaches to data analysis should have 1) strong prediction accuracy, 2) straightforward model interpretability and explainability, 3) high stability/robustness, and 4) fast learning capability using fewer training data points since in most practical cases, data can be expensive to obtain. However, such an ideal set of parameters is not easily obtained.

There exist numerous methods in the literature for analyzing and accurately learning a predictive model from a complex data set (Bishop, 2006). Model complexity (both in terms of number of tuned parameters and interpretability) of these algorithms spans a wide range, with relatively simple models like k-Nearest Neighbors (k-NN) at one end of the spectrum to state-of-the-art machine learning algorithms like Deep Belief Networks (DBN) on the other (Figure 3). Moving from left to right in Figure 3, the models generally improve in their ability to characterize the underlying relationships in the input corpus of data, and thus represent advances in achieving high prediction accuracies. However, the increased accuracy often comes at a cost of complex network architecture and high dimensional operating spaces, making it hard to communicate the representations learned by the algorithm in a format easily understood by a human interpreter. This restricts their usage in tasks where interpretability of the learning model and understanding of the underlying patterns in the data are important outcomes, such as for system design. Also, the large number of hyper-parameters (like number of hidden layers, sparsity regularization, learning rate, etc.) that need to be manually tuned or require significant knowledge of structure and operation of these machine learning models additionally limits their usefulness to a machine learning layperson.

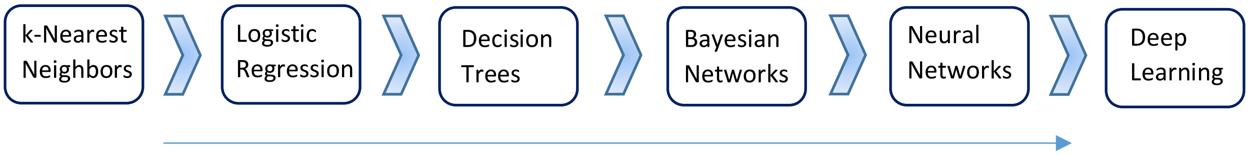


Figure 3: Complexity map of common machine learning models

On the other hand, less complex classification models such as decision trees and k-means clustering may provide insights that are more interpretable. Indeed, their popularity is highlighted by the fact that such clustering ML methods have dedicated packages and support available in R (open source programming language and software environment for statistical computing) and other programming languages such as MATLAB. However, it is not clear what it means to be an interpretable machine learning algorithm or what the tradeoff is between algorithm predictive ability and interpretability.

For example, there exist many clustering machine learning methods (e.g., hierarchical clustering and partitioned clustering (see Xu & Wunsch, 2005 for a review)), and the main advantage these algorithms have over the more complex machine learning models like deep belief networks (DBNs) is that the metrics used for computing clusters are easier to understand than the complex layered network structure of DBNs. However, while such clustering machine learning approaches may be more interpretable, most clustering approaches are highly sensitive to noise, which may result in poor prediction accuracy. Although there exist many feature selection and feature extraction techniques (e.g., Tang et al., 2014) to reduce the model sensitivity to noise, using them often transforms original input data into a new feature space (such as with principal component analysis). This transformation changes the contextual meaning of the features being used, making it difficult to use the model output to draw design inferences.

Consider the pedestrian experiment discussed previously; it is clear how age or conscientiousness are related to crossing decisions, in that as both go up, crossing times become more conservative. This relationship is harder to understand when it is a weighted linear combination of demographic variables that relate to particular crossing behaviors. Thus, it is difficult to find a single solution that includes all the characteristics of an ideal machine learning model, i.e. strong prediction accuracy, clear model interpretability, high stability/robustness and fast learning capability using fewer training data points.

To illustrate the challenge of selecting an appropriate machine learning algorithm for contextual cue analysis, the pedestrian dataset was tested with three machine learning approaches. Specifically, decision trees were used to classify pedestrians' decision times based on demographic traits and crossing positions (cross-walkers vs. jay-walkers) to better understand how the various displays on the car (Figure 2) impacted crossing behavior for different types of participants. Three different popular clustering algorithms for creating a decision tree were tested and compared: 1) Fast and Frugal Trees (FFT); 2) Classification and Regression Trees (CART); and 3) Evolutionary Trees (EvTree).

These three classification approaches differ widely in terms of model complexity (increasing from FFT to EvTree). Table 1 summarizes the strengths and limitations of each of the three approaches. As can be seen in Table 1, the differences in the way the classification trees are constructed and the relationships are learned create unique characteristics for each algorithm. While a designer might blindly apply one of these methods due to familiarity or ignorance of other approaches, these unique characteristics may impact the structure or interpretation of the resultant model.

As outlined in Table 1, it was not clear which one would yield the most useful results. Typically, people tend to favor models with a strong predictive accuracy. However, in many practical machine learning problems, prediction accuracy alone fails to provide any direct evidence to elucidate the stability and goodness of the trained model. In order to link ML to design, the results should be stable, i.e., regardless of which sample of the dataset was used to train the model, the results should generally be similar.

Using each of the three decision tree algorithms in Table 1, a classifier was trained on a random sample of the data (training data) and each element of the remaining data sample (testing data) was classified using the trained model. Each of the classifiers attempted to group the participants in the pedestrian experiment by their personality characteristics as a function of decision times to cross in order to understand the impact of the different displays, which was the primary design question for this experiment.

Table 1: Summary of the advantages & disadvantages of popular decision tree classification algorithms

Classification Algorithm	Advantages	Disadvantages
Fast and Frugal Trees (Gigerenzer & Todd, 1999; Luan et al., 2011; Bryant, 2002) Efficient and simple heuristic for classification tasks, inspired by human reasoning	1. Computationally fast, compared to all the above mentioned decision tree algorithms 2. Resultant decision trees are robust and less susceptible to over-fitting	1. Do not use all possible cues and do not integrate information while building decision trees 2. Since the heuristic computes no utility or probability to quantify the goodness of a branch split, it may lead to non-optimal splits
CART (Therneau & Atkinson, 1997; Breiman et al., 1984; Timofeev, 2004; Kim & Loh, 2011; Esmeir & Markovitch, 2007) Most commonly used classification method using GINI index as the splitting rule for building decision trees	1. No underlying assumptions about the nature of the observations, for example, to be independent and identically distributed 2. Results are invariant to the monotone transformation of the predictor variables, for example squaring a variable won't change the structure of the decision tree 3. Resultant decision trees are not sensitive to outliers	1. Possibility of non-optimal splits when learning a problem with strong interdependency among the predictor variables 2. Unstable decision trees; small variations in the training data set can lead to different tree structures
Evolutionary Trees (Grubinger, 2014; Deb, 2011; Barros et al., 2012) A globally optimal classification tree built using an evolutionary algorithm	1. Best suited for problems where multiple (locally optimal) solutions are needed; cases where the best solution may not always be realizable 2. Useful for problems with a huge search space, e.g. finding optimal decision trees which is NP-hard (Zantema & Bodlaender, 2000)	1. Computationally expensive and large memory requirements 2. Random nature of the algorithm can yield different tree structures with the same evaluation function value 3. Large number of parameters (crossover probability, mutation rate, no. of generations, etc.) that need to be manually tuned, mostly by trial-and-error approach.

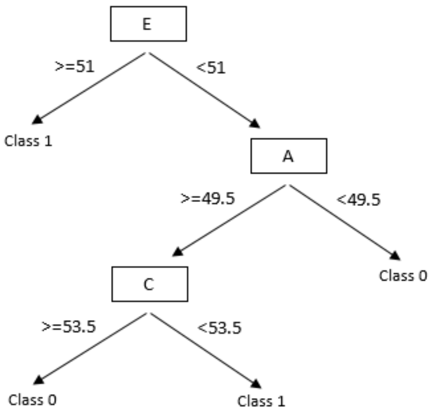
Figure 4 illustrates representative trees from the three different classification methods. Each method classifies a pedestrian into two classes, such that *Class 0* represents a cluster of pedestrians for whom the display on the car (Figure 2) did not matter and *Class 1* represents a cluster of pedestrians who leveraged the information from the car's display while making a crossing decision. The designations of E, O, C, and A in Figure 4 stand for extraversion, openness, conscientiousness and agreeableness which are the dominant personality traits of those participating in the experiment.

The three decision tree models vary in terms of tree structure and the type of variables used to cluster the data. For instance, the decision tree formed using FFTs shows that just simply being above the 51% (median) threshold on the extraversion personality scale predicted the use of one of the displays in Figure 2, a finding not revealed by the hypothesis-driven ANOVA. The CART approach further subdivided those people between *agreeableness and openness* into *Class 1 (*people who depended on the display). In comparison, the evolutionary tree tended to group participants similarly to the FFT and classified participants that relied on the display as primarily *extraverted* (a score >=49 or >= 53), but this is difficult for some to understand, as it appears that the *openness* root node is the primary relationship.

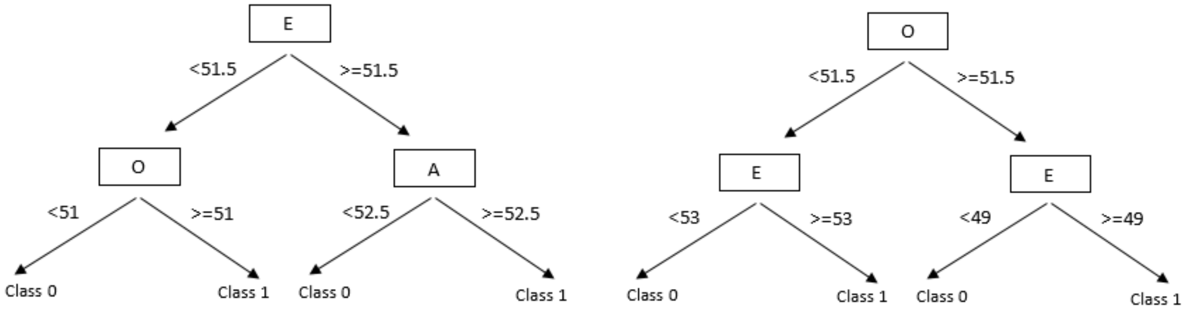
This process of training the model was repeated 1000 times for each classifier. Prediction accuracy was calculated by averaging the number of correct predictions at each iteration for every classifier (Figure 5). This process was carried out with 95% and 65% of the data set as the training samples (so 5% and 35% of the data was used for test samples). These two different splits were selected to study the stability of the classification algorithms (Figure 6). Only the first three nodes of the trained decision tree models were used to find optimal tree depth to minimize overfitting.

Observing the results in Figures 4, 5, and 6, highlight the following points:

1. The classification algorithms tend to cluster the data differently (Figure 4), making it difficult to consistently draw conclusions.



a) FFT



b) CART **c) EvTree**

Figure 4: Three clustering approaches produce different outcomes that are not in agreement

2. While CART tends to outperform the other two approaches in prediction accuracy (for 95% training (Figure 5)), it is less stable compared to FFT, as variables used at the first three levels of a decision tree to cluster the data are different for the 95% and 65% training cases (Figure 6).
3. The FFT approach could be seen as a more robust decision model because of the relatively higher prediction accuracy and stable clusters in scenarios where training is limited. However, from Figures 5 and 6, we can observe that for large training sets, CART tends to outperform the other approaches in terms of prediction accuracy and model stability. However, its model

becomes less robust with decreases in the number of training data points, although its predictive accuracy is still relatively high.

Also, the difference in the way these algorithms work can cause difficulty in interpretation of the results. For instance, the way the FFT method trains a decision tree makes it impossible to compare the clusters of multiple output classes together in one single representation. The FFT approach requires that a designer adopt a one-vs-all strategy, which means comparing between a large number of varied decision tree representations to account for all the class labels before arriving at a conclusion. For this specific example, three times as many decision trees had to be created for the FFT approach as compared to the CART approach, with an added intermediate classification interpretation step that introduces ambiguity in the results.

Given these results, we elected to continue the analysis with CART as it was the best algorithm in terms of our four criteria of 1) strong prediction accuracy, 2) straightforward model interpretability and explainability, 3) high stability/robustness, and 4) fast (enough) learning capability.

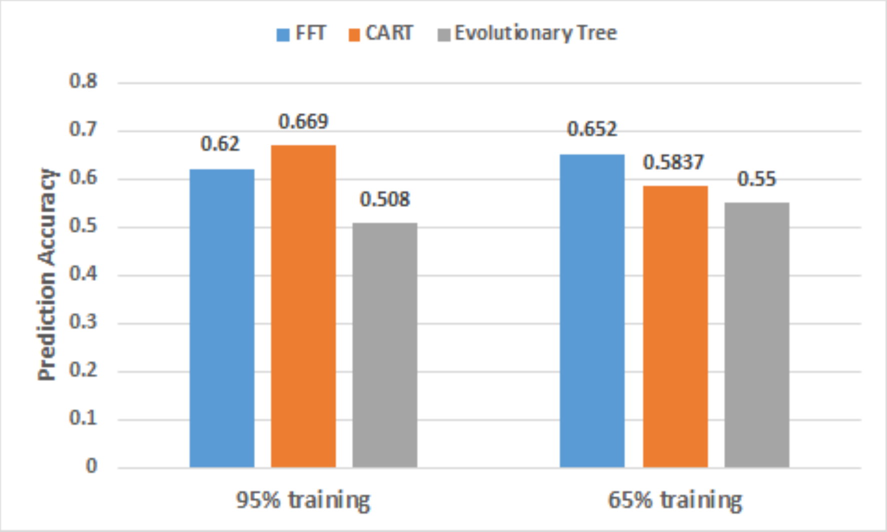


Figure 5: Average performance of different classifiers on the testing data set as measured by the prediction accuracy

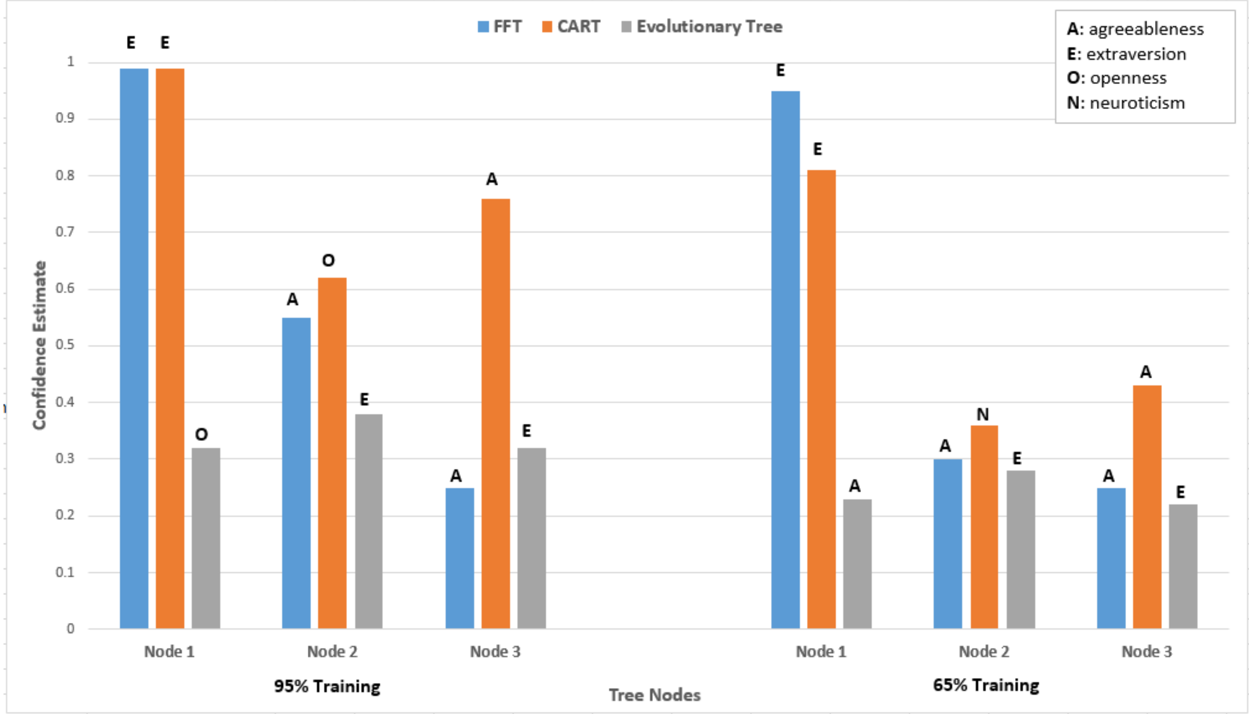
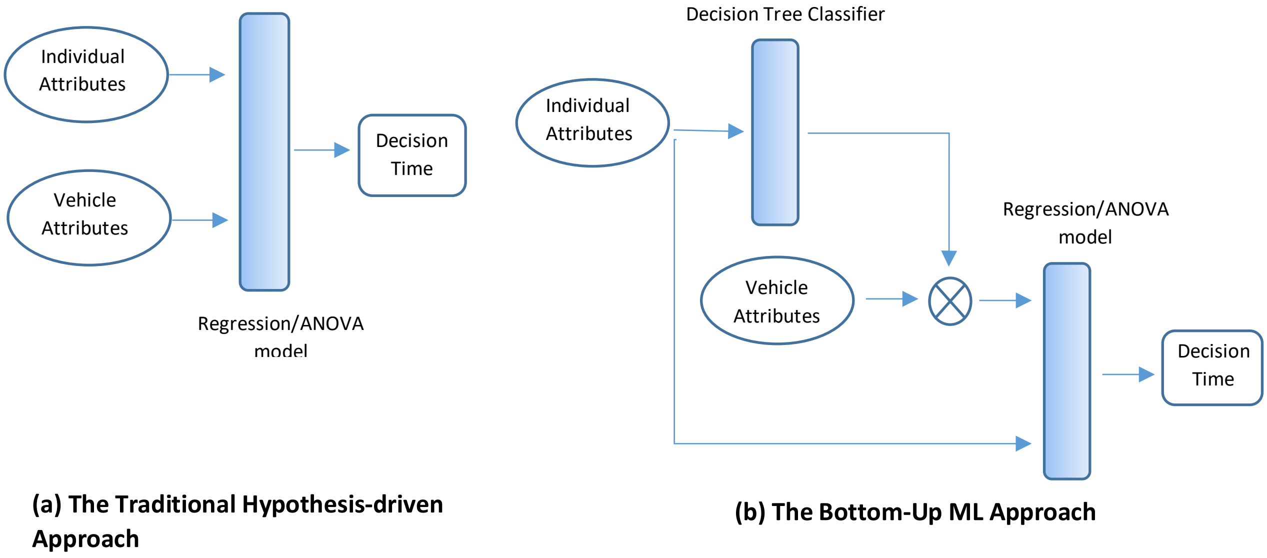


Figure 6: Measure of clustering algorithm stability. The X-axis represents the number of decision tree splitting nodes, with node 1 representing the root node. The Y-axis represents the confidence estimate of using a particular demographic variable (N, E, O, A) for clustering.

Understanding Contextual Cues from ML Analyses

As discussed previously, the traditional approach using top-down, hypothesis-driven experimental methods via an ANOVA led to the conclusion that neither the displays nor the speed of the car (the vehicle attributes) had any global effect on decision times and the only individual attributes that were statistically significant were age and conscientiousness (Figure 7a). However, we can frame the problem differently using a ML approach in that a bottom-up, data driven approach can be taken to first determine which segments of the population are most affected by the designs in question, and then develop a hypothesis-driven statistical model, e.g., a within-subjects ANOVA, on those clusters (Figure 7b).



(a) The Traditional Hypothesis-driven Approach

(b) The Bottom-Up ML Approach

Figure 7: Different approaches to data analysis: a) Treating independent variables as factors & performing inferential omnibus tests, b) Clustering people based on their individual attributes like personality, age and crossing position (cross walker vs. jay walker) in terms of which vehicle attributes they focused more on (display, speed or direction of the approaching car) and then performing statistical tests on these groups

This data-driven approach to the problem (Figure 7b) provides us with the flexibility to account for individual differences, which is an important contextual cue, for the people in our analysis, as opposed to aggregating the data across an expected population. In typical hypothesis-driven statistical analyses, individual differences are treated as uncontrollable variability (Trumbo et al., 2011). The use of blocked designs, covariance analyses, and other related pre and post hoc tools attempt to partition and minimize the effect of individual differences but doing so potentially causes researchers to lose important and useful information (Revelle, 1993; Davies et al., 2013). We hypothesize that by using ML to preprocess the data, we can actually identify and leverage individual differences to account for one source of context.

Using the multi-stage approach depicted in Figure 7b to reanalyze the pedestrian experiment results described previously with CART, the new results showed that some pedestrians do leverage the information from their surroundings including the external display, as opposed to just relying on legacy behaviors suggested by the traditional approach (Figure 7a). Of the original 55 subjects, 42% of participants using CART predominantly relied on the displays more than any other factor such as speed of the car or which side of the road they were on.

Given the novelty of self-driving cars and the fact that people have not had much exposure to new forms of vehicle-to-pedestrian communication, the demographic profiles of such early adopters is of great interest to car companies. Moreover, unlike in the traditional hypothesis-driven approach, the behaviors of this group of potential early technology adopters could give important insight to designers. For example, using the new population of early adopters identified through CART, the analysis as depicted in Figure 7b revealed that those who elected to cross in front of a vehicle had the fastest decision times if they used the Information display (which provided the vehicle's current speed, Figure 2c), followed by those using the advice display (Figures 2a & 2b), and then those with no display at all. Use of the information display for this group resulted in statistically faster decision times (3.33s) as compared to the next fastest time for the advice displays (3.44s, p= 0.040). This translates into an extra 4 ft., on average, in terms of distance away from an oncoming car, which has practical significance as well.

CONCLUSION

It is crucial when designing autonomous technologies to consider carefully how such systems can effectively interact with both operators and other relevant stakeholders, particularly in safety critical systems like self-driving cars and manufacturing and surgical robots. Such systems typically generate significant amounts of data, but it is not clear how industry designers can account for context and leverage analytic tools like machine learning to gain insight from this data into the actual use of intended designs or the influence from external, potentially problematic cues.

Machine learning has been used extensively for modeling user choice preferences, but little attention has been paid to how to use such techniques to gain new design insights into user behaviors, particularly in terms of understanding contextual cues, or to connect user behaviors to system performance. We propose that given the large amounts of data that such autonomous systems typically generate, machine learning could be useful if such algorithms were accurate, stable, learned relationships relatively fast, and were interpretable in a design context and explainable.

A case study was presented that highlighted the difficulties in selecting the best algorithm for the contextual cue analysis. Determining that CART was the best algorithm for this analysis, we demonstrated that applying machine learning techniques to the design data analysis can lead to interesting and potentially useful results that are very different from traditional hypothesis-driven statistical experimental designs. We are not suggesting that machine learning approaches should replace such scientific methods, but rather that they should be used to augment analyses.

Future related work should include determining what makes some ML algorithms better suited for design problems and what core characteristics define such utility. Moreover, given long-standing problems with people understanding probabilistic reasoning algorithms (Tversky and Kahneman, 1974), are there representations that make some ML algorithms more interpretable and explainable for industry users? Explainability of ML techniques is likely a multidimensional construct, and a future area of inquiry should be describing how and why various ML approaches may be more or less useful in the design context.

Acknowledgements

This research was supported by a grant from the National Science Foundation (NSF) (No. 1548417) to Duke University. The technical monitor was Jordan Berg. The views and opinions expressed are those of the authors and do not necessarily reflect the views of the NSF.

References

Amershi, S., & Conati, C. (2007). *Unsupervised and Supervised Machine Learning in User Modeling for Intelligent Learning Environments* Paper presented at the IUI'07, Honolulu, HI.

Andre, A. D., & Wickens, C. D. (1995, October 1995). When Users Want What's NOT Best for Them. *Ergonomics In Design, 1995,* 10-14.

Bainbridge, L. (1987). Ironies of Automation. In J. Rasmussen, K. Duncan, & J. Leplat (Eds.), *New Technology and Human Error*. New York: Wiley.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Berlin: Springer.

Boussemart, Y., Cummings, M. L., Las Fargeas, J., & Roy, N. (2011). Supervised vs. Unsupervised Learning for Operator State Modeling in Unmanned Vehicle Settings. *Journal of Aerospace Computing, Information, and Communication, 8*(11), 71-85.

Breiman, L. (2001). Statistical Modeling: The Two Cultures *Statistical Science, 16*, 199-231.

Clamann, M., Aubert, M., & Cummings, M. L. (2017). *Evaluation of Vehicle-to-Pedestrian Communication Displays for Autonomous Vehicles*. Paper presented at the The Transportation Research Board (TRB) 96th Annual Meeting Washington DC.

Cockburn, A., Karlson, A., & Bederson, B. B. (2009). A review of overview+detail, zooming, and focus+context interfaces. *ACM Computing Surveys, 41*(1).

Crowe, S. (Producer). (2016). Tesla Autopilot Causes 2 More Accidents. Retrieved from www.roboticstrends.com/article/tesla_autopilot_causes_2_more_accidents

Cummings, M. L. (2004). *Automation Bias in Intelligent Time Critical Decision Support Systems.* Paper presented at the AIAA 3rd Intelligent Systems Conference, Chicago.

Daily, M., Medasani, S., Behringer, R., & Trivedi, M. (2017). Self-Driving Cars. *Computer Vision and Image Understanding, 50*(12), 18-23.

DARPA. (2016). *Explainable Artificial Intelligence (XAI).* (DARPA-BAA--16-53). Washington DC: Department of Defense.

Feary, M., Billman, D., Chen, X., Howes, A., Lewis, R., Sherry, L., & Singh, S. (2013). Linking Context to Evaluation in the Design of Safety Critical Interfaces *Human-Computer Interaction. Human-Centred Design Approaches, Methods, Tools, and Environments* (pp. 193-202). Berlin: Springer.

Franke, J. L., Daniels, J. J., & McFarlane, D. C. (2002). *Recovering Context After Interruption.* Paper presented at the 24th Annual Meeting of the Cognitive Science Society, Fairfax, VA.

Gross, B. (2013). Google's Self Driving Car Gathers Nearly 1 GB/Sec. Retrieved from https://www.linkedin.com/pulse/20130502024505-9947747-google-s-self-driving-car-gathers-nearly-1-gb-per-second

Guerlain, S., Jamieson, G. A., Bullemer, P., & Blair, R. (2002). *The MPC Elucidator: a case study in the design for human-automation interaction.* Paper presented at the IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans.

Huang, X., Oviatt, S., & Lunsford, R. (2006). Combining User Modeling and Machine Learning to Predict Users' Multimodal Integration Patterns In S. Renals, S. Bengio, & J. Fiscus (Eds.), *Machine Learning for Multimodal Interaction* (Vol. 4299, pp. 50-62). Berlin: Springer-Verlag.

Johnson, C. L. (2014). Context and Machine Learning. In P. Brézillon & A. J. Gonzalez (Eds.), *Context in Computing* (Vol. 2, pp. 113-126).

Kraut, R. E., & Lewis, S. H. (1982). Person perception and self-awareness: Knowledge of influences on one's own Judgments. *Journal of Personality and Social Psychology, 42*(3), 448–460.

Lankenau, J. B. a. A. (2002). *A Rigorous View of Mode Confusion.* Paper presented at the SafeComp, Bremen, Germany.

Lee, J. D., & Morgan, J. (1994). *Identifying Clumsy Automation at the Macro Level: Development of a Tool to Estimate Ship Staffing Requirements.* Paper presented at the Human Factors and Ergonomics Society Annual Meeting Nashville.

Lesch, M. F., Powell, W. R., Horrey, W. J., & Wogalter, M. S. (2013). The use of contextual cues to improve warning symbol comprehension: making the connection for older adults. *Ergonomics, 56*(8), 1264–1279.

Marcus, G. (2018). Deep Learning: A Critical Appraisal. *arXiv*.

Nielsen, J. (1993). *Usability engineering*. Cambridge, MA: Academic Press

Nielsen, J. (2005). Putting A/B Testing in Its Place. Retrieved from https://www.nngroup.com/articles/putting-ab-testing-in-its-place/

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychology Review, 84*, 231–259.

Norman, D. A. (1983). Design rules based on analysis of human error. *Communications of the ACM, 26*(4), 254-258.

Norman, D. A. (1986). Cognitive Engineering. In D. A. Norman & S. W. Draper (Eds.), *User Centered System Design: New Perspectives on Human Computer Interaction* (pp. 31-61). Hillsdale, NJ: Lawrence Erlbaum.

Norman, D. A. (1988). *The Design of Everyday Things*. New York: Doubleday.

Pentland, A., & Liu, A. (1999). Modeling and prediction of human behavior. *Neural Computation, 11*, 229-242.

Ruskin, K. J., Stiegler, M. P., Park, K., Guffey, P., Kurup, V., & Chidester, T. (2013). Threat and error management for anesthesiologists: a predictive risk taxonomy. *Current Opinion in Anaesthesiology, 26*(6), 707-713.

Schneiderman, B. (1987). *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Reading, MA: Addison-Wesley.

Srinivasan, V. (2016). Context, Language, and Reasoning in AI: Three Key Challenges. *MIT Technology Review*.

Stimpson, A., & Cummings, M. (2014). Assessing Intervention Timing in Computer-Based Education Using Machine Learning Algorithms. *IEEE Open Access, 2*, 78-87.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science, 185*(4157), 1124-1131.

Vakil, S. S., & Hansman, R. J. (2002). Approaches to mitigating complexity-driven issues in commercial autoflight systems. *Reliability Engineering and System Safety, 75*, 133-145.

Webb, G. I., Pazzani, M. J., & Billsus, D. (2001). Machine Learning for User Modeling. *User Modeling and User-Adapted Interaction, 11*, 19-29.

Wei, Q. (2014). *A Context-Aware System to Communicate Urgency Cues to Health Care Workers.* Norwegian University of Science and Technology, Trondheim, Norway.

Wenxin, L., & Kekang, H. (2008). *Effects of Contextual Cues and Support Requirements of Multimedia Animation on Children's English Learning*. Paper presented at the Education Technology and Training, Shanghai.

Wickens, C. D., & Hollands, J. G. (2000). *Engineering Psychology and Human Performance* (3rd ed.). Upper Saddle River, N.J.: Prentice Hall.

Wiener, E. L., & Nagel, D. C. (1988). *Human Factors In Aviation*. San Diego, CA: Academic Press.

Wilson, T. D., & Nisbett, R. E. (1978). The accuracy of verbal reports about the effects of stimuli on evaluations and behavior. *Social Psychology, 41*(2), 118–131