Interruptibility Prediction for Ubiquitous Systems: Conventions and New Directions from a Growing Field

Liam D Turner, Stuart M Allen, Roger M Whitaker
Cardiff School of Computer Science & Informatics, Cardiff University
Cardiff, United Kingdom
{TurnerL9, AllenSM, WhitakerRM}@cardiff.ac.uk

ABSTRACT

When should a machine attempt to communicate with a user? This is a historical problem that has been studied since the rise of personal computing. More recently, the emergence of pervasive technologies such as the smartphone have extended the problem to be ever-present in our daily lives, opening up new opportunities for context awareness through data collection and reasoning. Complementary to this there has been increasing interest in techniques to intelligently synchronise interruptions with human behaviour and cognition. However, it is increasingly challenging to categorise new developments, which are often scenario specific or scope a problem with particular unique features. In this paper we present a meta-analysis of this area, decomposing and comparing historical and recent works that seek to understand and predict how users will perceive and respond to interruptions. In doing so we identify research gaps, questions and opportunities that characterise this important emerging field for pervasive technology.

Author Keywords

Interruptibility; Ubiquitous computing; Context-aware computing; Meta-analysis.

ACM Classification Keywords

H.1.2 User/Machine Systems: Software psychology

INTRODUCTION

Assessing another person's interruptibility prior to interaction with them is a natural human behaviour [24, 48] that is generally easily handled by the human brain. However, creating such capability in the context of a machine, so that there is harmonious synchronicity with human behaviour, is a significant challenge that has important ramifications for the demands placed upon a user. Historically, interruptibility has been studied in static task-oriented environments such as offices, using desktop computers (e.g., [20, 12, 26, 41]), or in controlled laboratory simulations (e.g., [16, 4, 39]).

UbiComp '15, September 7–11, 2015, Osaka, Japan.

http://dx.doi.org/10.1145/2750858.2807514

The introduction of new pervasive sources of interruption from ubiquitous technologies, in particular mobile devices, has increased the scope of the problem in spatial and temporal dimensions, subsequently impacting our daily lives [43, 46, 56]. These technologies utilise auditory, visual or haptic cues to either inform the user of available information (e.g. notifications) or attempt to immediately shift attention (e.g. a phone call). Consistent with this, interest has arisen from a wide range of disciplines including: psychology [38], HCI [37], and artificial intelligence systems [49, 9, 44], as well as diverse application areas including medical [48] and safety [32] related areas. This motivates examining the area in a unified way to reflect on the existing approaches and identify areas to further explore.

The primary contribution of this paper is a meta-analysis of the recent literature that provides a decomposition of the various conventions adopted for interruptibility research. Additionally, we propose a set of research questions from our meta-analysis that intend to stimulate potential research directions around the technological issues for inferring interruptibility; as well as social, privacy, and ethical concerns.

Objective and organisation of the paper

Broadly speaking, reviews and analysis of interruptibility studies have involved two distinct approaches. On the one hand, interruptibility has been encompassed within the concepts and visions of wider attention-aware systems [45, 49, 40, 54]. On the other hand, works have evaluated specific conventions relevant to interruptibility, such as the contextual features adopted [19]; whether to use experience sampling methods (ESM) [33] ; or human labelling by a third party [3]. In this paper we focus on approaches that:

- provide approaches for intelligent interruptions;
- test the interruption process in different scenarios, and gain feedback from the user.

In doing so, this paper provides an analysis and classification of intelligent interruption approaches in the context of ubiquitous computing. We believe that this is timely: the fast moving nature of the subject, and it's porous boundaries, mean that it is challenging to characterise new contributions and compare them against previous work.

In addressing this area, we assume that the relationship between the human and technology is such that technology ex-

1. Scenario Selection	2. Data Collection	3. Prediction
Decide on the: - interruptions used; - interruption environment; - intended objective	Decide on how to represent interruptions and response behaviour Collect data and extract feature vectors, including choosing the : - data traces to sample; - feature variable extraction process Label the extracted feature vectors Aggregate each instance to form a dataset	Perform pre-processing Build predictive models, including choosing the: - training method; (e.g. offline or online) - training data (e.g. personalised subset) Evaluate the predictive performance

Table 1. The typical linear paradigm of interruptibility studies, including subcomponents.

ists to augment and synchronise the individual's behaviour. Thus inappropriate interruption has a human cost (e.g. annoyance or cognitive burden), as does the lack of a legitimate interruption (e.g. opportunity cost). This standpoint is consistent with the supportive role of intelligent technology and motivates accurate interruptibility prediction. Interruptions in the right context have been argued to augment some task-oriented environments [23] or even provide productivity stimulus when self initiated [28].

We organise our analysis by first exploring the overarching conventions in which interruptibility is defined. We then explore how interruptibility research is undertaken by following the typical linear paradigm of: defining a scenario; data collection; and building predictive models (Table 1). Within each, we explore the design choices, assumptions, and implementation practices used. Additionally, the extent of interoperability and generalisation of technical approaches, beyond the context in which they are introduced is also discussed. Finally, within each of these areas we highlight emergent trends and issues, and propose open research questions (RQ) as potential directions going forward.

Overarching approach towards assessing interruptibility

Inferring interruptibility concerns identifying whether introducing a stimulus that the user may choose to act upon is suitable, typically decided *in situ*. Thus, to minimise disruption and maximise timely response rates, interruptions should ideally occur at the most convenient or opportune moments, where the disruption caused to an individual is minimised. However there are degrees of freedom within this, as highlighted by Ho and Intille [19]. In particular their 2005 survey reports at least 8 definitions of interruptibility and 11 measures that impact the perceived burden of an interruption, including functional activities, social context, historical patterns of behaviour and emotional state.

More generally, we suggest that the definition of what it means to be interruptible is fragmented across the literature, with 3 broad categories:

- *physiological ability* to switch focus;
- *cognitive affect* on task performance;
- *user sentiment* towards the interruption.

The *physiological ability* to switch focus involves assessing the cognitive workload of the user at the time of interruption, and their capacity to receive it. At the very lowest level the effect of mental workload on the user can be assessed using EEG [35] or pupil size events [2, 4], although achieving this outside of controlled conditions is currently not a practical basis for measurement.

The *cognitive affect* on task performance addresses the ability or overhead to switch from an existing task to an interruption and then re-engage back to previous task. This has typically been adopted in task-oriented environments through identifying breakpoints where disruption is minimised (e.g., [25, 39]). A common metric used is the elapsed time to regain focus after the interruption, commonly referred to as *resumption lag*, measured through software events (e.g., [25, 1, 39, 26]).

User sentiment captures the current emotional state in reaction to the interruption. This often involves more subjective metrics captured on a Likert scale using self reports (ESM) (e.g., [47, 43]). Some studies attempt to distinguish this concept from physical interruptibility, by determining this as *receptiveness* to an interruption [19, 10].

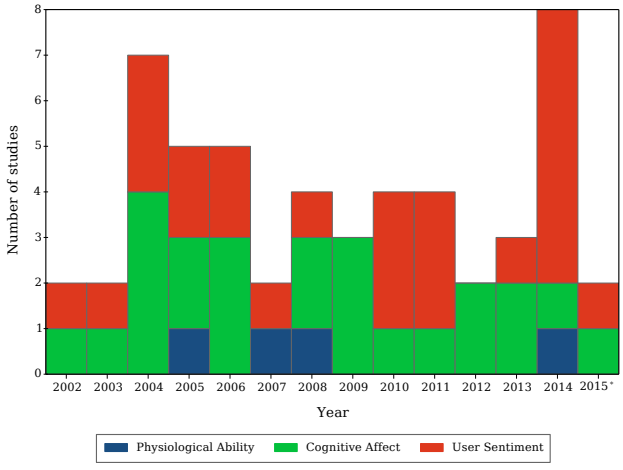

Figure 1. The distribution of interruptibility definition categories over time. * Meta-analysis was conducted before the end of 2015.

		Interruptibility Category			Experiment Environment			Data Collection			Labelling Strategy		
Study	**Year**	**PA**	**CE**	**US**	**OE**	**EITW**	**IITW**	**ESM**	**RS**	**SS**	**ESM**	**IOB**	**PCL**
[39]	2002		X		X				X			X	
[23]	2002			X	X			X			X		
[20]	2003		X		X				X		X		X
[24]	2003			X	X					X	X		X
[41]	2004		X			X		X	X		X		
[22]	2004			X	X				X		X		
[1]	2004		X	X	X			X	X		X	X	
[13]	2004				X				X		X		
[6]	2004		X		X				X				
[18]	2004		X		X				X			X	
[30]	2004			X	X				X		X		
[14]	2005		X		X				X			X	
[2]	2005	X			X				X			X	
[12]	2005			X	X					X	X		X
[21]	2005		X		X				X			X	
[19]	2005			X		X			X		X		
[25]	2006		X		X				X			X	
[49]	2006		X										
[5]	2006		X	X	X			X	X		X	X	
[31]	2006			X	X				X		X		
[35]	2007	X				X			X		X		X
[59]	2007			X		X		X			X		
[26]	2008		X	X	X				X		X	X	
[4]	2008	X	X		X				X			X	
[52]	2009		X										
[7]	2009		X										
[28]	2009		X		X			X					X
[10]	2010			X		X			X		X		
[27]	2010		X			X			X			X	X
[60]	2010			X					X		X		
[15]	2010			X		X		X			X		
[56]	2011			X	X				X		X		
[58]	2011		X		X				X		X		
[11]	2011			X		X			X			X	
[50]	2011			X		X			X			X	X
[34]	2012		X			X			X			X	
[29]	2012		X		X				X			X	
[40]	2013			X									
[54]	2013		X										
[16]	2013		X		X			X	X		X	X	
[57]	2014			X		X			X				
[46]	2014			X			X		X			X	
[47]	2014			X			X		X		X		
[55]	2014			X			X		X			X	
[53]	2014	X		X		X		X	X			X	
[43]	2014			X		X		X	X		X		
[8]	2014		X		X				X			X	
[42]	2015			X		X		X			X		
[32]	2015		X		X				X			X	

Table 2. A decomposition of the approaches used across studies, sorted ascending by year. PA=Physiological Ability, CE=Cognitive Effect, US=User Sentiment, OE=Observed Environment, EITW=Explicit "In-The-Wild", IITW=Implicit "In-The-Wild', ESM=Experience Sampling Methods, RS=Real Sensors, SS = Simulated Sensors, IOB=Implicit Observations of Behaviour, PCL=Post-Collection Labelling.

In conducting our meta-analysis, we've grouped recent works under these categories in Table 2, and visualised the disparity across works over time in Figure 1. We note that some studies consider multiple categories and others do not define an explicit definition of interruptibility; in these cases we have made our best judgement from the information provided. Overall, this supports previous claims that comparing studies is a difficult task [54].

It could be argued that this fragmentation is due to these definitions having different relevance for different scenarios. For

example, user sentiment is likely to be less relevant to nurses working in an emergency facility [48], while office environments are more relevant to sentiment and the cognitive affect on workload. Nevertheless, we observe a lack of standardised definitions of interruptibility, which creates additional barriers when comparing and building from relevant works.

DIMENSION 1: SCENARIOS FOR INTERRUPTIBILITY

The first dimension of interruptibility studies is defining the scenario. At it's highest level this captures the scope, by defining a *channel of interruption* (such as smartphone notifications), *an environmental consideration* (which addresses the physical context in which the interruption is studied), and the *objective for the study*.

In general, studies typically use a single channel for interruptions. This ranges from audio recordings (e.g., [13, 12]); to messaging communications (e.g., instant messaging [46, 16] or email [27]); to tasks in other PC application windows (e.g., [14]); to phone calls (e.g., [11, 55]); and to smartphone notifications (e.g., [43, 47, 51]). In reality, our daily lives involve multiple devices that can interact with us in more than one way. Additionally, these devices may have multiple means of interaction, they may be restricted by place or time, and multiple devices can exist at the same time. Exploring how interruptibility can be affected by different channels (i.e., *how* as well as *when* to interrupt) has been a relatively unexplored area, which leads to the following an open research question:

(RQ1) How can different channels of interruption (and potentially devices) be used in combination and to the best effect?

Sarter [54] reviews interruption management in a multimodal context, highlighting approaches that have been developed for different sensory channels. In particular, presenting the user's involvement and decision making in the interruption management process, as well as highlighting the performance costs of interruptions and proposing empirically based recommendations for modality choices given a range of scenarios. However more empirical work in this area is needed, particularly comparisons where the same interruption content is used.

Experiment environments have ranged from all moments of daily life (e.g., through a personal smartphone [43, 57]) through to a more specific focus, such as those with high social costs (e.g., [17]) or where task disruption is likely to occur, such as in offices (e.g., [12, 41]). More generally, the environments are either *controlled* or *in-the-wild*, as classified in Table 3. Controlled environments have traditionally involved a laboratory setting, providing close observation of behaviour and a restricted decision space within which activities are undertaken (e.g., [26, 29, 1]). However we argue that while more natural behaviour can be assessed, in some cases office settings may also fall into this category, such as when a third party observer is present (e.g., [28]) or when cameras are used (e.g., [12, 31]). A visualisation of the distribution of these works (from Table 2) is shown in Figure 2. There is a clear recent increase in experimenting *in-the-wild*, we suggest that this likely due to the spatial and temporal freedom

Type	Definition
Controlled environment	The experiment takes place in a static laboratory setting, involving simulations of activities and interruptions. Users are typically compensated for their time, but not always.
Explicit in-the-wild	The experiment takes place *in situ* around the daily live of participants. However, the user is continually aware of the experiment. The participants are typically incentivised through compensations for their time, but not always.
Implicit in-the-wild	The experiment takes place *in situ* around the daily lives of participants. The experiment is often embedded through other features that the participant finds useful, providing more natural incentive.

Table 3. Types of experiment environments used.

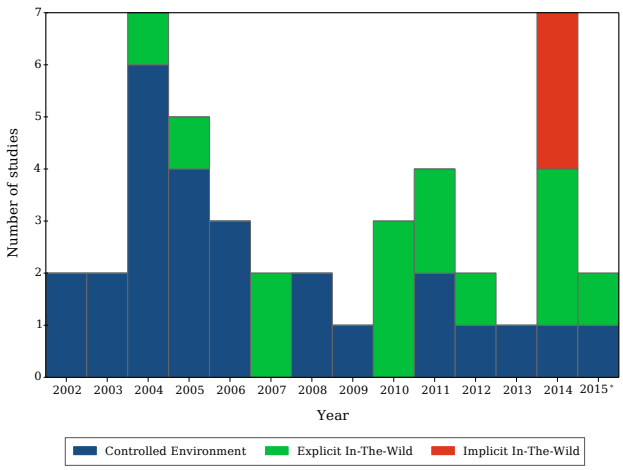

Figure 2. The distribution of experiment strategies over time. * Meta-analysis was conducted before the end of 2015.

that ubiquitous technologies such as the smartphone have enabled.

The objective for a study dictates to what extent different priorities are considered for making or assessing interruptions. For example, some papers have considered classifying all moments as either interruptible or not (e.g., [11, 47, 43]), while others have considered finding breakpoints within or between activities for interruptions to occur (e.g., [26, 58, 25]). There are also instances with specific focus, such as predicting the timeliness instant messages being read (e.g., [46]).

Overall, we note that the scenarios for studying interruptibility are heavily domain and interruption specific. The problem is that the choices made here have a profound effect on the later stages (e.g., what data is collected) and ultimately intelligent interruption systems. This creates uncertainty in assessing the interoperability for other scenarios, which could require costly implementation and testing to determine. Therefore, another open research question remains in whether a one-size-fits-all framework can be achieved or whether we are limited to grouping studies based on similar scenarios:

(RQ2) Given the diversity of potential scenarios, when are generalised and interoperable solutions for interruptibility sufficient, and when are domain specific solutions necessary?

We believe relatively little progress has been made on generalised approaches for interruptibility. Whilst some works attempt to generalise specific channels of interruption (e.g., smartphone notifications [47]), possible unification of interruptibility approaches for flexibility across different scenarios remains an ongoing area to explore.

DIMENSION 2: DATA COLLECTION

Data collection is a fundamental requirement; however there are considerable degrees of freedom in what data is collected, when, and how.

Representing context and interruption behaviour

Typically, interruptions are represented by a set (or vector) of variables that capture the context at a given moment, and a label representing either some categorisation of interruptibility or event dictating interruptibility (e.g., whether it is a cognitive breakpoint). In the case of categorising interruptibility, a typical convention has been to use a binary state (e.g., [53]) or to represent the degree of interruptibility on a scale (e.g., [56, 43]). This representation of each interruption attempt provides the basis for statistical analysis and machine learning, and sets the requirements for data collection.

This simplifies datasets into a set of cases where each states: given a context, the user was interruptible (or not). However this causes a reliance on the user completing a labelling process in response to the interruption (e.g., filling in a survey on their interruptibility [43, 47]). In reality, a user may be interruptible but not enough to complete that labelling process, or they may find doing so undesirable. It may also be the case that not all information is available to the user until they begin to respond (e.g. Android notifications). This could result in responses that are started but then abandoned, where arguably some degree of interruptibility is shown. This risks these cases being classified the same as those where no response was started, i.e. the user wasn't physically interrupted or they weren't interruptible enough to switch focus.

Investigations into the importance of incomplete responses has received little attention, in which we propose the following research question:

(RQ3) Can including the extent of a response to an interruption provide additional semantic value for inferring the user's attentiveness towards it?

Across the literature there is a foundation of key works introducing relevant concepts. McFarlane and Latorella show that the act of interrupting and responding is a decision process [36] and Pejovic and Musolesi discuss the concept of user attentiveness being a subcomponent of interruptibility [43]. We note that the exact decision process and ways in which a response can be abandoned is likely to be scenario dependent, which may be further constrained by the technical viability to collect this behaviour. However, there has been little empirical investigation into the impact these responses have on prediction accuracy if classed as either interruptible or not interruptible, or whether these cases should be classed separately.

Type	Data Traces
Context	Smartphone Sensors: e.g., hardware sensors [57, 32, 46, 47, 53, 43, 11, 50] or software APIs [34, 57, 46, 55, 11, 8, 50]
	Physiological sensors: e.g., physical state [32, 53] or activity [53, 30, 19]
	Environmental sensors: e.g., sound or motion in a room [41, 13, 24, 6, 20, 21] or car [32]
	Software events: e.g., active windows, keyboard and mouse activity [25, 26, 29, 41, 13, 58, 39, 18, 6, 20, 22, 21, 14]
	Calendar schedules [56, 57, 20]
	Temporal logs: e.g., of user actions [29, 34, 46, 21]
	Spatial logs: e.g., GPS [56, 57, 55, 53, 11, 50] or connections to antennas [41, 47, 55, 43, 22]
Latent	Self reports: e.g., experience sampling [10, 42, 41, 59, 53, 43, 23, 26, 30, 19] or post-experiment surveys [1, 16, 28]
	Qualitative feedback: e.g., post-interviews [10, 23]
	Third party observer reports: e.g., in situ observation [28] or video annotations [24, 30, 12]
	Physiological sensors: e.g., mental state or workload [35, 2, 53, 4]

Table 4. A categorisation of the commonly used data traces.

Representing the current context

A key design consideration is the choice of what data to capture to represent the current context. This involves a process of collecting from raw data traces and extracting feature variables. Depending on the study, a top-down approach may be taken where the variables are decided first. Alternatively, a bottom up approach may be taken where the exact features are decided after collection of a range of raw data traces.

Data traces

The aim of collecting data is to capture signals from which a representation of the current context can be made. Table 4 details the types of data traces collected from, and classifies them as either contextual or latent sources. These can loosely be described as what is currently happening and what the user feels respectively. Ideally, data should be as rich as possible, and could span multiple modalities, however resource constraints and scenario environments typically dictate a subset of these being used (as shown in Table 4) and could involve different means of collecting the data (as shown in Table 2).

Advancements in ubiquitous sensing technology (such as the smartphone) has seen the adoption of real sensors (e.g., [47, 6, 43]) rather than simulated sensors (e.g., [24, 12]) in recent years (Figure 3) as a means to collect these data traces. We suggest that this is likely due to the need for post-experiment annotations no longer being a constraint. However there is still disagreement over whether smartphone sensors should be used [43, 47, 57] or not, due to accuracy and reliability issues [56, 33, 53] and resource requirements [55]. However, the personal relationship between these technologies and their

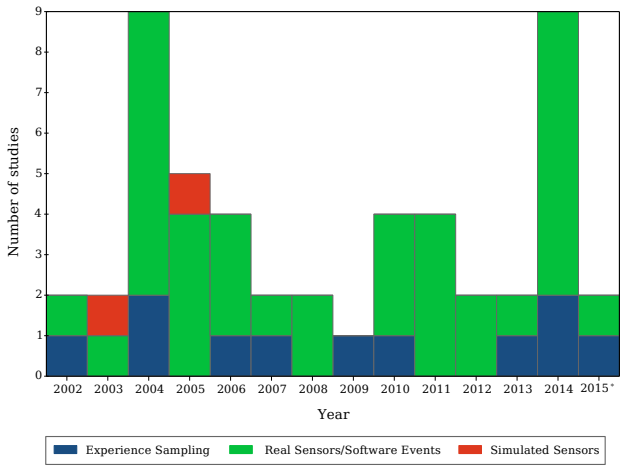

Figure 3. The distribution of data collection strategies over time. * Meta-analysis was conducted before the end of 2015.

user has been argued to allow more "ecologically valid data" [38], rather than having peripheral devices, such as cameras (e.g., [24, 12]) or wearable accelerometers (e.g., [19, 30]). Nevertheless, the persistence of ESM suggests that the overall stability and limitations in observing latent variables remains an issue.

Additionally, we note from the literature that generally speaking, a single or very limited number of devices are used in combination. More recently, we observe a trend in the consolidation of sensing devices, such as only using a smartphone (e.g., [43, 47, 11]). The emergence of communication enabled pervasive technologies in the environment (i.e. the internet of things) and upon the person (i.e. smart wearables), could augment existing data traces (shown in Table 4) (e.g., a light sensor in a room might be more suitable than a smartphone equivalent in a pocket), or extend the possibilities of what contextual data is possible to collect - leading to the question:

(RQ4) How can emerging sensor-equipped ubiquitous technologies (such as wearables) improve sampling accuracy and reduce collection and processing complexity in-the-wild?

Tapping into these technologies could be promising for future research, with their presence becoming more natural and accepted (like the smartphone), rather than the presence of foreign peripheral devices introduced just for experiments.

Alongside sensor-based collection, human feedback has been a key source of data traces; either from a third party observer (e.g., [28]) or by the participant themselves (e.g., [43, 23]). Participant feedback has typically been collected through experience sampling methods (ESM) (e.g., [59, 42]), often involving the user answering questions to a survey *in situ*. The benefits of this include being generally applicable, having a low cost overhead in terms of technical resources, and allowing the collection of latent variables which aren't easily observable by readily available sensors [33]. However, the use of ESM for interruptibility research specifically has been controversial due to the additional interruption cost it places on the user [11], as well as the questionable accuracy and consistency of human quantification [41, 42].

Type	Features
User Features	Pupil size events [2, 4], EEG events [35], emotion [53, 43, 15], learning style [57], personality [57], time until next calendar event [22, 56]
Environment Features	Coarse location [59, 42, 47, 53, 43], fine location [57, 56, 41, 23], other people present [24, 12, 43], states e.g. door open/closed [12, 13], cell tower id [55], wifi ssid [55], nearby bluetooth [43], wireless signals [22], smartphone ringer state [46, 11], smartphone screen covered [46, 47, 11], smartphone orientation [47], ambient noise [11, 6]
Interruption Features	Content e.g. text or phone number [55, 10, 50], task complexity [16], number of queued interruptions [46], time between interruptions [20]
User and Environment Features	Time of day [12, 57, 56, 41, 22, 46, 55, 47, 50], day of week [57, 22, 46, 55, 53, 50], month [55], user is in conversation [59, 22, 12, 53, 22, 21], user's current activity [12, 41, 57, 53, 43, 23, 28, 30, 50], user is present [12, 6, 20, 19, 24], software events [46, 22, 41, 34, 58, 39, 18, 8, 6, 28, 13, 20, 21, 14], unusual environment to be in [42], frustration level [57, 1, 26], stress [53], level of annoyance [5] respiration [53], ambient sound [41, 22, 20], car movement [32], human motions [32, 19], smartphone motions or acceleration [47, 11], PC active and inactive time [22]
User and Interruption Features	Social relation [59, 16, 50, 15], interruption frequency [16], content desirability [42], perceived mental effort [1, 16], perceived task performance [16, 1], resumption lag [25, 1, 34, 39, 26], perceived timeliness of delivery [42], number of primary task errors [29, 5], primary task duration [1, 34, 22, 5], elapsed time to switch to interruption [46, 53, 21, 26], primary task complexity [57, 16], interruption timestamp [55], interruption duration [34, 1, 53, 5], perceived time pressure [1], previous or next task cue presented [29], elapsed time before user reaction [18], influence from social contexts [15]

Table 5. A categorisation of the common variables used for modelling, extracted from data traces.

Feature Variables

After collecting data traces, feature variables are extracted from the raw data to create a flat structure representing the current context. A common first step is to apply smoothing techniques to the data, in order to remove noise (e.g., [35]). From the meta-analysis conducted, we can see no evidence of widely adopted conventions within interruptibility studies - potentially due to the use of different data traces and hardware. Whilst technically challenging, this is an issue to consider going forward as this can affect the conclusions of the statistical analysis and machine learning.

Generally speaking, these variables can be categorised as representing either the: user, environment, interruption, or the relationships between these. A previous survey by Ho and Intille [19] detailed 11 measures/variables that have previously been considered to influence interruptibility. However, due to the volume and breadth of studies since, we have extended their observations and detailed the commonly used variables in Table 5. It should be noted that the variables included here were identified where they were either explicitly stated

or a reasonable level of confidence could be assumed. Whilst some features are scenario dependent, we observed great differences across works in the features used, with only a few reoccurring often. Again this supports that comparing and building from interruptibility works is challenging [54].

Assessing the suitability of feature variables is a common practice when attempting to reduce the footprint of predictive models in relation to classifier performance. However, given the complexity associated with collecting and transforming data traces to feature variables, it would be useful to quantify the resource cost that this brings. Choosing appropriate and technically feasible data sources is common at the design phase, but reflecting on the cost post-collection has received little attention. When operating in environments with highly constrained resources, such as the smartphone, this could bring valuable design considerations for future studies and applications:

(RQ5) Can the utility of potentially influential variables be standardised by considering the trade-off between accuracy and sampling / processing complexities?

Several works have touched on this within wider domains. For example, Lathia et al explore the issues relating to smartphone sensor sampling stability [33]. However, this is not common practice for interruptibility studies. A standardised means of quantifying the cost of retrieving individual feature variables on specific hardware or empirical evidence of the difficulties of doing so would be valuable for future design considerations.

Labelling interruptions

Labelling instances of interruptibility is necessary for classification purposes, however accomplishing this can be problematic. Two distinctly different approaches have been dominant in this area, explicit or implicit labelling. Explicit labelling typically involves direct labelling by the user through self reporting (ESM) (e.g., [19, 12, 15, 11, 43]). Likewise to collecting data traces, whether a user can accurately and consistently quantify their interruptibility, both in real time (e.g., [19, 12]) or retrospectively (e.g., [42]), has been brought into question [55]. Alternatively, implicit labelling involves observing user actions and making deductions (e.g., [11, 46, 55]). For a smartphone, this may be observing whether a phone call is answered or a notification is dismissed.

Figure 4 visualises the adoption of these labelling practices over time. We find that retrospective labelling has not been widely used in recent years, likely due to technological advances enabling participant feedback *in situ*. Interestingly, the debate of using ESM or implicit behaviour observation is reflected in the consistent use of both techniques over time.

Datasets

Datasets collected for the study of interruptibility have predominantly involved a small number of subjects (up to approximately 20 participants) as seen in [13, 55, 43] and up to approximately 100 as seen in [50, 16, 46, 47], with larger analysis of thousands of users being an uncommon and relatively recent occurrence, as seen in [51, 34]. It could be

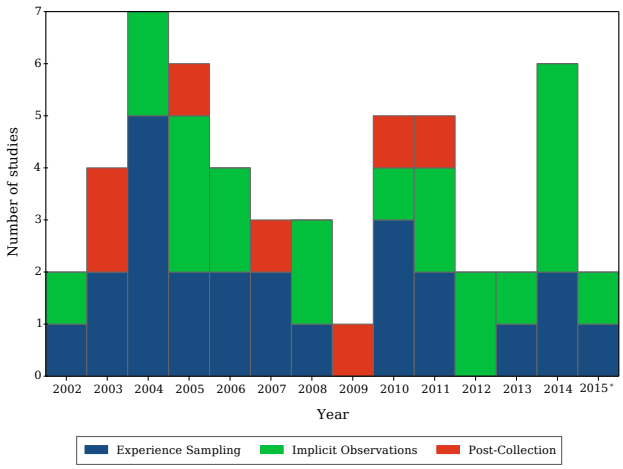

Figure 4. The distribution of labelling strategies over time. * Meta-analysis was conducted before the end of 2015.

assumed that more data from more users for longer is better, however there has also been investigations into reducing dataset size in later stages, to balance the footprint of predictive models with accuracy [13, 11]. Establishing suitability guidelines for dataset size and diversity has received little attention. However, we note that there has been support for the importance of longitudinal data, to observe interruptibility habits over time (e.g., [35, 43, 29, 55]).

Additionally, there has been little attention towards the scalability and sustainability of the architecture to collect datasets. Historically, forming a dataset involved manually retrieving the data from each participant (e.g., [19]), whereas the introduction of technologies such as the smartphone has enabled a more autonomous client-server model, supporting *in-the-wild* studies (e.g., [47]). With data traces potentially becoming more diverse and integrated into our daily lives (as noted by RQ4), this raises a key research question:

(RQ6) What architectural barriers remain in enabling the collection, storage, and processing of detailed sensor data and interruptibility behaviour at scale? More specifically, what roles should sensors, personal devices and servers play to minimise connectivity and processing bottlenecks?

Several approaches have highlighted architectural frameworks encompassing wider intelligent interruption systems (e.g., [57, 40]). However, we note from the meta-analysis that empirical evidence of feasibility at scale is lacking. This could be supported from works in other wider areas, including: cloud computing, high performance computing and network aspects of other areas within ubiquitous computing.

Extending from this is the social, ethical and privacy standpoint for this architecture and the resulting datasets, in which we propose the following research question:

(RQ7) What consent and anonymisation measures are appropriate for applications and researchers to know how interruptible someone is, and how does this balance with the potential for bias from the knowledge of behaviour monitoring?

This area has also received little attention but is arguably fundamental to the viability of interruptibility research for real-world applications. With this in mind, there are currently are also no conventions to provide "open data", impeding reproducibility of results, or consideration of benchmark data sets for alternative analysis methods by other researchers, which would be invaluable for future studies.

In addition, obtaining quality data requires user engagement, which in turn requires appropriate incentivisation. However, if incentives cause deviations from natural behaviour they can adversely affect a study and its conclusions [38]. The balance of informed consent and behavioural bias extends beyond interruptibility into the wider research space of observing and learning from human behaviour [38]. Popular methods within interruptibility studies for addressing bias and incentivisation include: using monetary compensation (e.g., [14, 42, 25]); providing feedback and visualisations to the user; or implement an additional utility (e.g., mood diary features [47]).

The convention of experimenting *in-the-wild* (e.g., [43, 51, 35]) also addresses this bias to an extent by removing the locality limitations of a controlled experiment, promoting natural behaviour [38]. Ubiquitous technologies are enablers for this and it is becoming an increasingly popular approach (Figure 2), as it can also operate within the conventions that the user is already comfortable with, such as mobile applications (e.g., [47]). However this only mitigates some data quality issues. For example, in many cases participants in such studies are self-selecting, which can be challenging to control both the quantity and the quality of data.

DIMENSION 3: PREDICTION

We focus on prediction through machine learning as our final stage, explored through an examination of feature selection, classifier training, personalisation and evaluation. It should be noted that not all works study prediction; some simply apply statistical tests (e.g., [34, 29]) to determine whether certain factors correlate with interruptibility. While the conventions followed are relatively uniform, we note that studies unnecessarily hinder interoperability by not considering the boundaries of generalisation beyond the confines of the study.

Finding influential features

It is plausible to assume that some chosen feature variables may provide more predictive power than others. Commonly referred to as *feature selection*, this (optional) step aims to reduce the size of predictive models by investigating the impact of adding or removing each variable to the model. Common techniques for this include using a statistical correlation-based approach [12, 59] or a wrapper-based approach [12, 53], where subsets of features are evaluated to quantify their effect on classification performance. Our analysis showed that direct comparisons of these techniques are uncommon. However, Fogarty et al [13, 12] showed no significant difference between selection methods for accuracy in their study, but the fewer features typically selected in a wrapper-based approach was deemed favourable.

Feature ranking is another technique used to measure the influence of features; it is more abstract than the other methods by decoupling from a particular ranking measure. Fogarty et al [12] use an *information gain* metric, whilst Pielot et al [46] define their measure using the number of classifications which become incorrect after the feature is removed. Additionally, some works do not perform feature ranking, but do observe the common presence of features across different generated classification models [55].

Overall, these techniques balance accuracy with the number of features to reduce model size and mitigate issues such as *overfitting*. However the possibility of design constraints for devices such as low-end smartphones suggests that these shouldn't be the only metrics when aiming to reduce features. An assumption is often implicitly made that the features extracted are accurate and reliable, however there has been evidence that this may not to be the case [33, 38]. An extension of RQ5 would be to include how the quantified utility of a variable can be used with feature selection techniques.

From datasets to training sets

Predictive models are typically trained from a subset of the dataset, with the remainder used for testing the model. Several techniques have been adopted by interruptibility studies, the most common being cross-validation (e.g., [31, 13, 53, 43, 35]). This involves splitting the dataset into a training and a testing set multiple times (typically 10 folds) and using the mean performance, mitigating potential skewness from using a single training set. Less common methods of splitting training data from a dataset are also used. For example, Sarker et al [53] attempt to reduce the training data needs by creating additional representations of datasets using groups of cases at opposite polarities (in this case the 6 quickest and the 6 slowest responses). However, it is unclear whether this would be applicable beyond the confines of that individual experiment.

As this process is independent of the dataset size, a common objective has been to reduce the amount of data needed to train a model (e.g., [13, 11]). This reduces the overall complexity and improves viability for real-world applications [13] by reducing storage and processing requirements. In practice however, studies have had varying success with this practice. Fogarty et al [13], showed evidence of diminishing returns (using more than 40% of the original dataset) in the accuracy more training data brings. However, Fisher and Simmons [11] show clear fluctuations in the accuracy as more training data is considered, across several classifiers.

Whilst the typical focus has been on reducing the volume of training data, the usefulness of this is arguably limited beyond reducing the footprint of static models within the confines of that experiment. If another study wishes to apply the same approach, or for systems where data is captured over time, knowing what diversity and temporal representation the training data needs would be more useful. However this has received little attention, from this we propose the following:

(RQ8) How do training dataset characteristics such as feature diversity and temporal representation affect the diminishing returns of prediction performance?

Smith et al consider *concept drift* in their analysis [55], where the values for some features may only appear in the test data,

	Classifier	Works	Classifier	Works
Offline	Naive Bayes	[24, 13, 12, 11, 46, 55, 43, 59, 41, 14]	Support Vector Machines	[24, 12, 11, 46, 55, 53]
	(Derivatives of) Decision Trees	[24, 12, 11, 47]	Adaboost (w/ decision stumps)	[24, 12, 43]
	Bayesian Network	[20, 43, 22]	Logistic Regression	[50, 46]
	Random Forests	[46, 32]	(Derivatives of) Nearest Neighbour	[11, 55]
	Neural Networks	[47]	JRip	[47]
	RUSBoost	[55]	Genetic Programming	[55]
	Association Rule Learning	[55]	Adaptive Neuro Fuzzy Inference System	[57]
	Partial Least Squares	[16]		
Online	Naive Bayes	[43, 55]	Hoeffding Tree	[43]
	k-Nearest Neighbour	[55]	Support Vector Machines	[55]
	RUSBoost	[55]	Ozaboost	[43]

Table 6. An overview of classifier algorithms used for interruptibility prediction.

hindering the opportunity for an optimal model. However, further attention towards this issue is needed at individual study level, as this will have profound benefits for the viability for real-world systems.

Training online vs offline

Historically, interruptibility works have predominantly focused on offline learning (e.g., [59, 24]), where data is collected in advance and used to train a static model. A contributing factor to this has been the technical constraints from using distributed sensors without connectivity (e.g., [12]) and where there also may not be a means of processing data and training models *in situ*, or redelivering models autonomously between a server and client. However, we note that offline learning also has strengths, such as typically aggregating data from multiple subjects.

In contrast, the connectivity offered by the smartphone has allowed sensing, processing, decision-making systems, and feedback loops to either be entirely centralised [43] or split between the device and a server [57]. This allows the exploration of online learning approaches *in-the-wild*, where models are revised regularly as new data is collected. Typically, this approach has been used for personalised models [43, 57], where issues such as availability of initial data [43, 55] can be mitigated by improving models over time. Conclusions on which is better to use when is still arguably in its infancy, we therefore highlight the following ongoing research question:

(RQ9) When should intelligent interruption systems adopt online and offline learning, and what factors in the scenario and data collection influence this choice?

We note from the literature that the majority of works consider a single technique, with only a few recent studies directly comparing performance (e.g., Smith et al [55]) and others considering both individually in the analysis (e.g., Pejovic and Musolesi [43]). Further comparative evidence would be a useful contribution to the area, particularly using hardware capable of both (e.g. smartphones).

Personalisation vs composite models

The debate of aggregating user data to create composite models (e.g., [12, 47, 46, 53]) against personalising models for each user (e.g., [31, 50, 43]) often concludes in favour of personalisation. This is due to the variety of environments, activities, interruptions and preferences across users in their daily life [43, 55]. Personalised models have typically been associated with the adoption of online learning (e.g., [43, 55]), however there is an issue of having little training data initially.

Due to the fragmentation of works in terms of scenarios and features, it is hard to draw strong conclusions on which is better overall, or whether this is scenario dependent. Further to this, mixing personal and aggregated data as a hybrid approach has also been suggested [13]. In this case, aggregated data from other users could provide the initial model, which could then be removed as personalised data becomes available [13]. This could also mitigate against the issue of over-personalisation, instead of introducing randomness. As with offline and online learning, we note an absence of guidelines across the literature of when each is more appropriate, in which we propose:

(RQ10) Do personalised models mean better performance and how does this balance with increased complexity? Could a hybrid approach using personal and aggregated data reduce the training requirements for new users?

As machine learning involves the use of at least one of these approaches, all works highlighted in Table 6 provide a basis for addressing this research question. However, we note that the number of comparative works is limiting (e.g. Pejovic and Musolesi [43] compare learning from a combined set of unordered cases and personalised ordered cases). Additionally there has been little empirical investigation into a framework for facilitating a hybrid approach to training data, where any complexity and accuracy trade-offs could be improved upon.

Classification performance

The introduction of machine learning suites such as Weka [13, 12, 55, 43, 59] and MOA [43] have enabled straightforward analysis of machine learning classifiers and parameters. However the variations in the exact design choices adopted presents challenges when comparing works (as evidenced in Table 6), worsened further by the other varieties in scenarios and data collection practices. The overarching problem is that it is difficult to assess the likelihood of the predictive model being interoperable to other studies or applications.

Across studies, several classifiers have been used (Table 6), with the most common domains being: tree, rule, function or Bayesian based algorithms. A typical convention has been to experiment with multiple classifiers and choose which has the best performance, sometimes by using statistic tests (e.g., [12]). However, as with feature selection, accuracy may not be the best metric for resource sensitive technologies. In these cases the computational complexity associated with generating and storing the model (or the connectivity requirements for sending the data to and from a server). Some works consider complexity when choosing classifiers (e.g., [50]), however this is not a widely adopted convention.

For evaluation, some studies compare performance against a baseline (e.g., [55, 59]) or human estimators (e.g., [24, 12, 56]). A popular method for creating a baseline has been classifying all moments as not interruptible (e.g., [13, 12, 47]) or interruptible (e.g., [59]). We suggest that this is likely motivated by different scenario objectives. Alternatively, Pejovic and Musolesi [43] create a bespoke baseline that "calculates the ratio of training set interruptions which resulted in a user reaction, and then in the simulator activates a notification with the probability that corresponds to that ratio". Whether a means of generalisation is achievable given the different objectives (e.g., avoiding unsuitable interruptions or exploiting possible opportunities) remains an ongoing area to explore.

In terms of reporting results, reporting confusion matrices is a common practice (e.g., [13, 53, 41]), or metrics calculated from them, such as precision and recall (e.g., [47, 43]), F-measure scores (e.g., [53]), or plotting true positives against false negatives (e.g., [35]). Less common metrics include Kappa statistics (e.g., [53]) and area under curve values (e.g., [47, 35]). Finally, more bespoke metrics have also been used, for example, Pielot et al [46] introduces a penalty system for misclassifications, using a higher cost when being non-interruptible is misclassified. We note from our analysis that the justification is typically (but not always) seemingly driven by machine learning conventions, rather than interruptibility objectives, where an incorrectly predicted suitable moment arguably has a greater negative impact than an incorrectly predicted unsuitable moment [24, 46, 55].

The overarching theme across these components is a difficulty to determine the suitability beyond the confines of the particular study. This raises questions of possible performance differences if a different feature selection or training method was used for example, or if the scenario and data collection practices changed. To enable more comparative work, an opportunity exists to construct a framework for evaluating and presenting classifier performance. We envisage an abstraction layer sitting below the convention of machine learning, in which we encourage the promotion of 4 components:

1. A means to benchmark classifier performance against either the most commonly used classifiers (Table 6) or a subset with similar computational complexity.

2. Using these results, discuss or show how the performance may change beyond the current investigating scenario, e.g. where resource sensitive technologies are used.

3. Performance should be compared against baseline conventions for unsuitable interruption cost and/or missed opportunity cost, using statistical tests.

4. Analyse how together, and individually, performance metrics (e.g., precision and recall) can be maximised in relation to interruption/opportunity costs.

As an extension of RQ2, this is a non-trivial problem, where the solution may not be a unified approach to conducting and evaluating interruptibility studies, but means in which the likely boundaries of interoperability can be better predicted.

CONCLUSIONS

The ability to perceive the interruptibility of another human being is an ability that has fundamental ramifications on our effectiveness to communicate. The introduction of pervasive technologies capable of interruption, such as the smartphone, has extended the impact of a machine's inability in our daily lives. As a result, research has focused on building towards intelligent systems for managing interruptibility. However, in doing so the specific scope in terms of: types of interruptions, environments and objectives has left the boundaries of the problem difficult to define [54].

The implications of this study is twofold, firstly, we conduct a meta-analysis of existing literature structured around the 3 linear stages that studies typically take: scenario selection, data collection, and predictive modelling - and the subcomponents within these. Whilst works have previously evaluated specific areas of relevance (e.g., [3, 33, 54]), we add to these by providing a holistic analysis of interruptibility research directions over time. We identify that not only is this an evolving research area, but there are several fundamental issues that require greater attention. Secondly, we propose 10 research questions towards the development of intelligent interruption systems. These include gaps to address in specific subcomponents, as well as wider trends, such as a lack of substantial exploration into the boundaries of generalisation.

REFERENCES

1. Adamczyk, P. D., and Bailey, B. P. If not now, when?: the effects of interruption at different moments within task execution. In *Proc. CHI'04*, ACM (2004), 271–278.

2. Adamczyk, P. D., Iqbal, S. T., and Bailey, B. P. A method, system, and tools for intelligent interruption management. In *Proc. TAMODIA'05*, ACM (2005), 123–126.

3. Avrahami, D., Fogarty, J., and Hudson, S. E. Biases in human estimation of interruptibility: effects and implications for practice. In *Proc. CHI'07*, ACM (2007), 50–60.

4. Bailey, B. P., and Iqbal, S. T. Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Transactions on Computer-Human Interaction (TOCHI) 14*, 4 (2008), 21.

5. Bailey, B. P., and Konstan, J. A. On the need for attention-aware systems: Measuring effects of

interruption on task performance, error rate, and affective state. *Computers in Human Behavior 22*, 4 (2006), 685–708.

6. Begole, J. B., Matsakis, N. E., and Tang, J. C. Lilsys: sensing unavailability. In *Proc. CSCW'04*, ACM (2004), 511–514.
7. Boehm-Davis, D. A., and Remington, R. Reducing the disruptive effects of interruption: A cognitive framework for analysing the costs and benefits of intervention strategies. *Accident Analysis & Prevention 41*, 5 (2009), 1124–1129.
8. Böhmer, M., Lander, C., Gehring, S., Brumby, D. P., and Krüger, A. Interrupted by a phone call: Exploring designs for lowering the impact of call notifications for smartphone users. In *Proc. CHI'14*, ACM (2014), 3045–3054.
9. Campbell, A., and Choudhury, T. From smart to cognitive phones. *Pervasive Computing, IEEE 11*, 3 (2012), 7–11.
10. Fischer, J. E., Yee, N., Bellotti, V., Good, N., Benford, S., and Greenhalgh, C. Effects of content and time of delivery on receptivity to mobile interruptions. In *Proc. MobileHCI'10*, ACM (2010), 103–112.
11. Fisher, R., and Simmons, R. Smartphone interruptibility using density-weighted uncertainty sampling with reinforcement learning. In *ICMLA'11*, vol. 1, IEEE (2011), 436–441.
12. Fogarty, J., Hudson, S. E., Atkeson, C. G., Avrahami, D., Forlizzi, J., Kiesler, S., Lee, J. C., and Yang, J. Predicting human interruptibility with sensors. *ACM Transactions on Computer-Human Interaction (TOCHI) 12*, 1 (2005), 119–146.
13. Fogarty, J., Hudson, S. E., and Lai, J. Examining the robustness of sensor-based statistical models of human interruptibility. In *Proc. CHI'04*, ACM (2004), 207–214.
14. Fogarty, J., Ko, A. J., Aung, H. H., Golden, E., Tang, K. P., and Hudson, S. E. Examining task engagement in sensor-based statistical models of human interruptibility. In *Proc. CHI'05*, ACM (2005), 331–340.
15. Grandhi, S., and Jones, Q. Technology-mediated interruption management. *International Journal of Human-Computer Studies 68*, 5 (2010), 288–306.
16. Gupta, A., Li, H., and Sharda, R. Should i send this message? understanding the impact of interruptions, social hierarchy and perceived task complexity on user performance and perceived workload. *Decision Support Systems 55*, 1 (2013), 135–145.
17. Harr, R., and Kaptelinin, V. Interrupting or not: exploring the effect of social context on interrupters' decision making. In *Proc. NordiCHI'12*, ACM (2012), 707–710.
18. Ho, C.-Y., Nikolic, M. I., Waters, M. J., and Sarter, N. B. Not now! supporting interruption management by indicating the modality and urgency of pending tasks. *Human Factors: The Journal of the Human Factors and Ergonomics Society 46*, 3 (2004), 399–409.
19. Ho, J., and Intille, S. S. Using context-aware computing to reduce the perceived burden of interruptions from mobile devices. In *Proc. CHI'05*, ACM (2005), 909–918.
20. Horvitz, E., and Apacible, J. Learning and reasoning about interruption. In *Proc. ICMI'03*, ACM (2003), 20–27.
21. Horvitz, E., Apacible, J., and Subramani, M. Balancing awareness and interruption: Investigation of notification deferral policies. In *User Modeling 2005*. Springer, 2005, 433–437.
22. Horvitz, E., Koch, P., and Apacible, J. Busybody: creating and fielding personalized models of the cost of interruption. In *Proc. CSCW'04*, ACM (2004), 507–510.
23. Hudson, J. M., Christensen, J., Kellogg, W. A., and Erickson, T. I'd be overwhelmed, but it's just one more thing to do: Availability and interruption in research management. In *Proc. CHI'02*, ACM (2002), 97–104.
24. Hudson, S., Fogarty, J., Atkeson, C., Avrahami, D., Forlizzi, J., Kiesler, S., Lee, J., and Yang, J. Predicting human interruptibility with sensors: a wizard of oz feasibility study. In *Proc. CHI'03*, ACM (2003), 257–264.
25. Iqbal, S. T., and Bailey, B. P. Leveraging characteristics of task structure to predict the cost of interruption. In *Proc. CHI'06*, ACM (2006), 741–750.
26. Iqbal, S. T., and Bailey, B. P. Effects of intelligent notification management on users and their tasks. In *Proc. CHI'08*, ACM (2008), 93–102.
27. Iqbal, S. T., and Horvitz, E. Notifications and awareness: a field study of alert usage and preferences. In *Proc. CSCW'10*, ACM (2010), 27–30.
28. Jin, J., and Dabbish, L. A. Self-interruption on the computer: a typology of discretionary task interleaving. In *Proc. CHI'09*, ACM (2009), 1799–1808.
29. Jones, S. A., Gould, S. J., and Cox, A. L. Snookered by an interruption?: use a cue. In *Proc. BCS-HCI'12*, British Computer Society (2012).
30. Kern, N., Antifakos, S., Schiele, B., and Schwaninger, A. A model for human interruptability: experimental evaluation and automatic estimation from wearable sensors. In *Proc. ISWC'04*, vol. 1, IEEE (2004), 158–165.
31. Kern, N., and Schiele, B. Towards personalized mobile interruptibility estimation. In *Location-and Context-Awareness*. Springer, 2006, 134–150.
32. Kim, S., Chun, J., and Dey, A. K. Sensors know when to interrupt you in the car: Detecting driver interruptibility through monitoring of peripheral interactions. In *Proc. CHI'15*, ACM (2015), 487–496.

33. Lathia, N., Rachuri, K. K., Mascolo, C., and Rentfrow, P. J. Contextual dissonance: Design bias in sensor-based experience sampling methods. In *Proc. UbiComp'13*, ACM (2013), 183–192.
34. Leiva, L., Böhmer, M., Gehring, S., and Krüger, A. Back to the app: the costs of mobile application interruptions. In *Proc. MobileHCI'12*, ACM (2012), 291–294.
35. Mathan, S., Whitlow, S., Dorneich, M., Ververs, P., and Davis, G. Neurophysiological estimation of interruptibility: Demonstrating feasibility in a field context. In *Proc. 4th International Conference of the Augmented Cognition Society* (2007).
36. McFarlane, D. Comparison of four primary methods for coordinating the interruption of people in human-computer interaction. *Human-Computer Interaction 17*, 1 (2002), 63–139.
37. McFarlane, D. C., and Latorella, K. A. The scope and importance of human interruption in human-computer interaction design. *Human-Computer Interaction 17*, 1 (2002), 1–61.
38. Miller, G. The smartphone psychology manifesto. *Perspectives on Psychological Science 7*, 3 (2012), 221–237.
39. Monk, C. A., Boehm-Davis, D. A., and Trafton, J. G. The attentional costs of interrupting task performance at various stages. In *Proc. HFES'02*, vol. 46, SAGE Publications (2002), 1824–1828.
40. Moran, S., and Fischer, J. E. Designing notifications for ubiquitous monitoring systems. In *Proc. PerCom'13 (PERCOM Workshops)*, IEEE (2013), 115–120.
41. Mühlenbrock, M., Brdiczka, O., Snowdon, D., and Meunier, J.-L. Learning to detect user activity and availability from a variety of sensor data. In *Proc. PerCom'04*, IEEE Computer Society (2004).
42. Patil, S., Hoyle, R., Schlegel, R., Kapadia, A., and Lee, A. J. Interrupt now or inform later?: Comparing immediate and delayed privacy feedback. In *Proc. CHI'15*, ACM (2015), 1415–1418.
43. Pejovic, V., and Musolesi, M. Interruptme: designing intelligent prompting mechanisms for pervasive applications. In *Proc. UbiComp'14*, ACM (2014), 897–908.
44. Pejovic, V., and Musolesi, M. Anticipatory mobile computing: A survey of the state of the art and research challenges. *ACM Comput. Surv. 47*, 3 (2015), 47:1–47:29.
45. Petersen, S. A., Cassens, J., Kofod-Petersen, A., and Divitini, M. To be or not to be aware: Reducing interruptions in pervasive awareness systems. In *Proc. UBICOMM'08*, IEEE (2008), 327–332.
46. Pielot, M., de Oliveira, R., Kwak, H., and Oliver, N. Didn't you see my message?: predicting attentiveness to mobile instant messages. In *Proc. CHI'14*, ACM (2014), 3319–3328.
47. Poppinga, B., Heuten, W., and Boll, S. Sensor-based identification of opportune moments for triggering notifications. *Pervasive Computing, IEEE 13*, 1 (2014), 22–29.
48. Rivera, A. J. A socio-technical systems approach to studying interruptions: Understanding the interrupter's perspective. *Applied ergonomics 45*, 3 (2014), 747–756.
49. Roda, C., and Thomas, J. Attention aware systems: Theories, applications, and research agenda. *Computers in Human Behavior 22*, 4 (2006), 557–587.
50. Rosenthal, S., Dey, A. K., and Veloso, M. Using decision-theoretic experience sampling to build personalized mobile phone interruption models. In *Pervasive Computing*. Springer, 2011, 170–187.
51. Sahami Shirazi, A., Henze, N., Dingler, T., Pielot, M., Weber, D., and Schmidt, A. Large-scale assessment of mobile notifications. In *Proc. CHI'14*, ACM (2014), 3055–3064.
52. Salvucci, D. D., Taatgen, N. A., and Borst, J. P. Toward a unified theory of the multitasking continuum: from concurrent performance to task switching, interruption, and resumption. In *Proc. CHI'09*, ACM (2009), 1819–1828.
53. Sarker, H., Sharmin, M., Ali, A. A., Rahman, M. M., Bari, R., Hossain, S. M., and Kumar, S. Assessing the availability of users to engage in just-in-time intervention in the natural environment. In *Proc. UbiComp'14*, ACM (2014), 909–920.
54. Sarter, N. Multimodal support for interruption management: Models, empirical findings, and design recommendations. *Proceedings of the IEEE 101*, 9 (2013), 2105–2112.
55. Smith, J., Lavygina, A., Ma, J., Russo, A., and Dulay, N. Learning to recognise disruptive smartphone notifications. In *Proc. MobileHCI'14*, ACM (2014), 121–124.
56. Stern, H., Pammer, V., and Lindstaedt, S. N. A preliminary study on interruptibility detection based on location and calendar information. *Proc. CoSDEO'11* (2011).
57. Sykes, E. R. A cloud-based interaction management system architecture for mobile devices. *Procedia Computer Science 34* (2014), 625–632.
58. Tanaka, T., and Fujita, K. Study of user interruptibility estimation based on focused application switching. In *Proc. CSCW'11*, ACM (2011), 721–724.
59. Ter Hofte, G. H. Xensible interruptions from your mobile phone. In *Proc. MobileHCI'07*, ACM (2007), 178–181.
60. Zulkernain, S., Madiraju, P., Ahamed, S. I., and Stamm, K. A mobile intelligent interruption management system. *J. UCS 16*, 15 (2010), 2060–2080.