Gaver, William, Bowers, John, Kerridge, Tobie, Boucher, Andy and Jarvis, Nadine

Anatomy of a failure: how we knew when our design went wrong, and what we learned from it

Anatomy of a Failure
How We Knew When Our Design Went Wrong, and What We Learned From It

William Gaver, John Bowers, Tobie Kerridge, Andy Boucher, Nadine Jarvis

Interaction Research Studio
Goldsmiths, University of London
London SE14 6NW, UK
initial.surname@gold.ac.uk

ABSTRACT
In this paper, we describe the failure of a novel sensor-based system intended to evoke user interpretation and appropriation in domestic settings. We contrast participants' interactions in this case study with those observed during more successful deployments to identify 'symptoms of failure' under four themes: engagement, reference, accommodation, and surprise and insight. These themes provide a set of sensitivities or orientations that may complement traditional task-based approaches to evaluation as well as the more open-ended ones we describe here. Our system showed symptoms of failure under each of these themes. We examine the reasons for this at three levels: problems particular to the specific design hypothesis; problems relevant for input-output mapping more generally; and problems in the design process we used. We conclude by noting that, although interpretive systems such as the one we describe here may succeed in a myriad of different ways, it is reassuring to know that they can also fail, and fail incontrovertibly, yet instructively.

Author Keywords
Ubiquitous computing, interpretation, home, failure

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
In this paper, we describe an interactive system that we designed, implemented, field tested, and ultimately concluded was unsuccessful.

Although it is commonly argued that failure is instructive, reports of failing designs are rare in the literature. To be sure, the reasons our system was unsuccessful cannot be known unequivocally, but the empirical evidence we gathered, coupled with our own reflections, does give us confidence in describing the shortcomings of the prototype we built. We discuss these shortcomings, and the lessons they might teach us, at three levels: problems particular to the specific design hypothesis, problems relevant for mapping input to output more generally, and problems in the design process we used in developing the system. Tracing how our system failed from these perspectives may provide useful insights into how to do better in the future.

CHI 2009, April 4–9, 2009, Boston, Massachusetts, USA.

What constitutes failure?
First, however, we describe the *symptoms* of failure – that is, how we knew that our system did not work. This is more problematic than might be supposed. Most systems reported in the HCI literature are designed to achieve predefined goals such as allowing some task to be completed or problem to be solved. This permits the establishment of criteria such as speed and accuracy against which system success can be measured. The system we report here, in contrast, embodies a style of design, and design research, in which human-machine interaction is seen as locally situated meaning making and the role of design as the provision of multilayered resources for this process [1, 4, 5, 7, 10, 12]. The open-endedness of this approach raises challenges for how systems should be evaluated, because what it means to succeed, and indeed the dimensions relevant for success, may vary widely depending on how people achieve a meaningful relationship with a given design. From this perspective, judging success or failure is not a trivial thing.

Approaches to evaluating interpretive systems such as the sort we describe here tend to focus on how to go about gathering suitable material for assessment, but avoid discussing how success or failure might be determined. For instance, Höök et al. [8] based their evaluation of a system on analysing the conversations that groups of people had on encountering it. Others seek alternatives to verbalised judgements to capture more intuitive and sensual aesthetic and emotional responses [9]. Finally, others advocate gathering multiple forms of evaluation from a variety of

perspectives, including those of 'cultural commentators' such as journalists or filmmakers [4]. Opening out evaluation to multiple voices and new forms of expression in these ways reflects the multiple interpretations afforded by the class of systems in which we are interested. On the other hand, these approaches can invite a kind of relativism from which it is difficult to draw firm conclusions.

In this paper, we propose features of user engagement as being reliably symptomatic of success or failure, based on the empirical work described here and in previous studies. We group these under four themes: engagement, reference, accommodation, and surprise and insight. As we shall see, this orientation to evaluation does not lead us to relativistic conclusions. On the contrary, an important goal here is to show how open-ended systems and strategies for their 'polyphonic' assessment need not be incompatible with making definitive assessments of success or failure that can help shape future design work.

The aspects of engagement we discuss here are not just relevant to open-ended systems. More traditional systems, which privilege particular tasks and interpretations, may also show the features of engagement we describe here. Indeed, these aspects may be crucial for assessing and understanding how the most utilitarian systems succeed and fail outside the laboratory, even though the more focused criteria typically used to test such systems may overshadow their consideration. From this point of view, our study of systems that do not have readily definable goals and criteria is useful in highlighting issues relevant for systems that do.

The rest of this paper is divided into three sections. In the first, we describe the system in question, including the ideas behind it, a previous instantiation, and the field trial of the current version. In the second, we describe the symptoms of success and failure that emerge from a comparison of the unsatisfactory experiences observed in this field study with more rewarding deployments of other systems in the past. In the third, we consider the reasons for the present prototype's failure and suggest lessons for future design-based research. Finally, we conclude with a discussion about the utility and generality of our findings.

THE HOME HEALTH MONITOR

The unsuccessful prototype we focus on here is called the Home Health Monitor, and is the second version of a previously reported system called the Home Health Horoscope [6]. We refer to the overall class as 'Home Health systems'.

The Home Health systems are intended to promote reflection about wellbeing in the home, where 'wellbeing' is defined loosely to refer to attributes such as 'sociability' or 'busy-ness' or 'disruption', that are discovered to be important based on interviews with the household itself. A number of sensors are deployed to measure states of the home which, based on a series of visits and interviews with household members, appear symptomatic of emotionally relevant activities (see Figure 1). The sensor data is processed to assess data trends and anomalies, and the results used by a rule-based system to create a vector-based representation of 'wellbeing' in which scores on a number of 'wellbeing metrics' (e.g. 'busy' or 'social') embody the system's inferences. This representation of wellbeing is in turn used to generate an output that displays the system's interpretation of domestic wellbeing to the household occupants.

For instance, we might design a sensor device to measure when a given door is open or shut because the home's occupants have informed us that it is only closed when household members want to avoid each other. The raw sensor data is processed to uncover attributes such as the total time the door is open or closed during the day, how often it is changed, or how early it is first moved. Rules compare the day's readings with trends found over the preceding days to determine whether they are unusually high or low, and map this to an increment or decrement of, e.g., the 'sociality' metric accordingly. The pattern of metric scores provides a representation of the home's wellbeing which is mapped to an output for users. In the first iteration [6], the system constructed 'horoscopes' from sentences culled from online examples and categorised according to the wellbeing metrics; in the system described here we tried three different forms of output.

From the users' point of view, then, Home Health systems are meant to provide an intriguing reflection of the household's 'mood' [2, 11]. More deeply, however, they were developed to test and demonstrate a broader conjecture about ubiquitous computing systems.

The 'User Appropriated Inference' Concept

The original notion of the Home Health systems emerged as one of a large number of sketch proposals produced in the early stages of a project on domestic technologies. The outcome of design-led research, the proposals were not intended to explicitly illustrate any particular theoretical hypotheses, but instead to explore potentially engaging design possibilities in a synthetic fashion [13]. Once the Home Health proposal was produced, however, we recognised that it embodied a particular design concept in an elegant way, and this became a large part of its appeal.

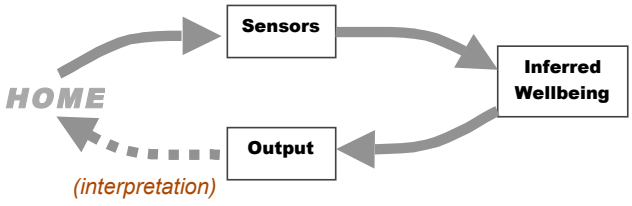

Figure 1: Basic system logic.

The foundation for this concept is the belief that so-called 'smart home' technologies, in which event inferencing algorithms from (e.g.) artificial intelligence research are used on data produced by ubiquitous computing sensors, are unlikely in principle to produce accurate accounts of the complex and idiosyncratic events and interpretations of home. The notion was that, despite this, the accounts produced by such systems may encourage and provide resources for people's own more accurate accounts, particularly if the automatically generated outputs are represented in such a way as to undermine system authority. In other words, if there is a continuum between effective randomness and total accuracy in systems' ability to represent the home, we believed we could locate a 'sweet spot' between the two in which systems might spur user interpretation of events in ways that would be based upon, but be more accurate than, the interpretations of the technical system itself (see Figure 2).

A number of interesting implications would follow if this *User Appropriated Inference* concept were valid. It would mean that ubiquitous computing systems would not have to build comprehensive representations in order to support user understanding. Instead, much like information visualisation software, the trick would be for such systems to provide information in a way that would support people's own pattern recognition abilities. This approach might help alleviate concerns for intrusiveness and invasion of privacy, since if system inferences are assumed to be inherently flawed, the emphasis should be on developing more evocative sensors and displays rather than more accurate ones. Moreover, accurate user inferences would depend on local knowledge, limiting the ability for outsiders to use system data in meaningful ways. Finally, an approach based on user appropriated inferences might generalise to a great many ubiquitous computing applications, including for instance systems to support aging in place or energy efficiency.

Thinking about the appeal of the Home Health systems quickly became synonymous with appreciating the User Appropriated Inference concept. As we shall see, the identification of a design with an abstract idea in this way turned out to be unfortunate for the system's development.

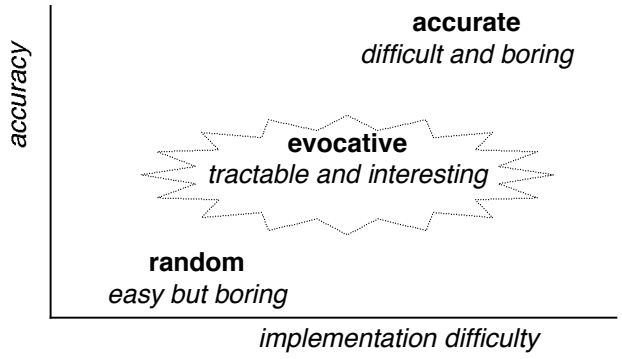

Figure 2: The basic hypothesis: partial accuracy may evoke user interpretation.

From Home Health Horoscope to Home Health Monitor

The system reported here grew out of the theoretical and design concerns outlined above, with particular features motivated by our experience with a first iteration.

This first iteration, the Home Health Horoscope, was developed in participation with a fairly large household in North London consisting of a nuclear family with children in their late teens and early twenties as well as a changing cast of partners, friends and lodgers who stayed with them for varying lengths of time. We studied how their routines manifested themselves in sense-able attributes of their household during occasional visits over more than a year, and developed a series of a dozen sensor devices and a set of about thirty rules specifically for their household. These rules determined wellbeing metrics relevant for their arrangements, and were used to generate 'horoscopes' automatically that were printed out once a day on a device in their home (see [6] for details).

The household lived with the resulting system for several months, during which time we assessed their experience using a combination of ethnographic observations and interviews, documentary film, and informal encounters occasioned by maintenance visits. Overall, the results were encouraging: we found that household members, and particularly our lead informant, engaged with the system continually throughout the deployment, regularly reading the horoscopes and relating them to ongoing activities. The horoscopes and overall system were the subject of many conversations within the household. Crucially, these discussions often centred not on whether the system understood the state and activities of the household accurately (e.g. 'the household is busy today)' but whether its interpretation of these affairs (e.g. 'you should slow down') was appropriate. In agreeing with the former while taking authority over the latter, the participants demonstrated the kind of relationship we had hoped to evoke.

The deployment was not an unmitigated success, however. The continual engagement with the system appeared motivated as much by questions about our research agenda as by interest in what the system was saying about the household. The outputs were seen as wrong quite often, to the extent that at least some participants speculated that the sensors might simply be fakes. Thus in the second implementation, we sought to build on the successes while resolving some of the limitations of the first. Here we describe the new features of the Home Health Monitor in terms of the shortcomings of the first prototype.

A Simpler Household

The original participants indicated that, while the system usually seemed to track state and activities adequately, it sometimes appeared to miss events or overinterpret the importance of some readings. In part this appears to reflect the difficulty of using only a few sensors to capture information about a large and complicated household. In addition, the size and construction of the original site meant that it was difficult to maintain radio contact between the sensors and the central server.

For the new implementation, then, we recruited a much simpler household consisting of a couple living in a single-story apartment in London. We met with them periodically over several months to familiarise ourselves with their routines and to consider how sensors might pick up relevant information about them. As we expected the routines turned out to be simpler, and the physical space more tractable, than in the first household.

'Readings' not Horoscopes

Despite the appeal of horoscopes as a culturally familiar genre that invites people to consider external interpretations of their lives, the first system uncovered several problems in their use. First, short sentences are difficult to automatically classify, which undermines some of the appeal of using found text. Second, extracts from horoscopes often imply particular contexts that may be irrelevant for a particular instance of a general category (i.e. they rarely comment on 'sociability' as an abstract concept, but instead as embodied in particular contexts such as going to parties or meeting strangers). Third, horoscopes can have undesirable cultural connotations either as a genre or in the particular approaches taken to writing them (e.g. our volunteers objected that the horoscopes we used were too didactic and suggested we emulate the style of a particular horoscope writer they liked).

In developing the Home Health Monitor, then, we decided to abandon horoscopes as an output style and instead to investigate different sorts of outputs. Initially we developed a series of 'readings' as output from the system. These were short sentences, often in the form of aphorisms that attempted to comment on household activities, sometimes in a mildly judgemental way, without unduly specifying particular contexts (e.g. 'work is a natural form of relaxation'). Later, as we realised that the system was failing to engage the volunteers, we experimented with two other forms of output. The first used photographs, sourced from the web and given to us by the volunteers from their personal collections, that we categorised according to the system's metrics. Along with these, we produced pie charts representing the scores on the day's well-being metrics.

Sensor Legibility

In the original system, we housed the sensor boards in small project boxes that were not intended to be inspected by the volunteers. We assumed that, because the focus of the system was on the horoscopes that embodied an integrated response to sensor readings, the sensor devices themselves would be of little interest to the volunteers (i.e. they would 'disappear'). This proved not to be true. The volunteers speculated about what exactly was being monitored throughout the trial. At times their conjectures became relatively suspicious, and they wondered whether we might be recording sound or images and even disassembled one of casings to inspect the sensor boards. These suspicions were short-lived, and the volunteers did not express significant enduring reservations to us about living with the system. Nonetheless it became clear that, far from 'disappearing', the sensor devices were objects of continual interest.

Rather than seek to make the new sensors less obtrusive, we focused instead on making them more *legible* – that is, on making the sensor readings that were being made clearer. First, integral to the designs were clear indications of the *sources of information* being monitored. For some, this involved wires leading to external pressure pads to be mounted under cushions. One included a cable to plug into the external audio output from a TV. Several included external light sensors shielded in such a way as to make their orientation clear (see Figure 3). Second, to the original sensor configuration (Crossbow System's MicaZ motes, piggybacked with MTS310 sensor boards) we added small displays that indicated the *number of events* the sensors had picked up. We hoped that these additional features would make the function and current activity of the sensors visible to volunteers, reassuring them about what was being monitored, and supporting them in reasoning about the overall operation of the system.

Transparent Hand-Off

Conjectures about the sensor devices were part of a wider pattern of speculation about the overall purpose of the system. In keeping with our normal practice, we had not told the volunteers about our intentions for the system to avoid biasing their interpretation of it. For the Home Health Horoscope, however, this lack of explanation prompted extensive questioning on the part of the volunteers about what the system might be doing and why. Hypotheses ranged from suspicions about spying, to conjectures that the horoscopes were merely a pretext for testing sensors in domestic environments, to accounts that came close to our own. This uncertainty seemed clearly to stem from the lack of explanation that we gave.

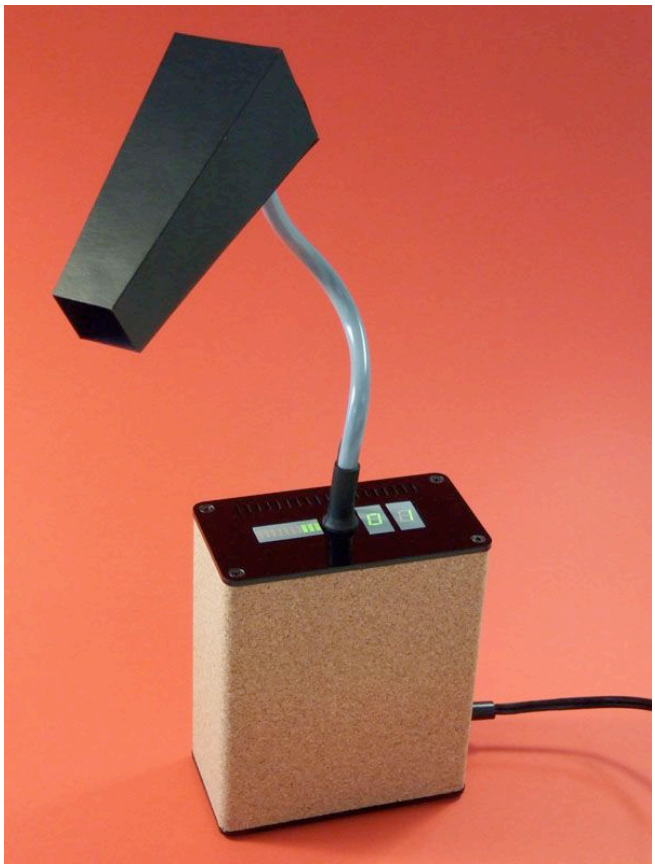

Figure 3: A sensor device designed for legibility.

One of the design changes we made for the new iteration, then, was to the process of deploying the system. Over the course of preparatory visits we made clear our interest in understanding how 'well-being' might become manifest in their home, and possibilities for using sensors to monitor this. When we installed the prototype in their home, we tried to balance clarity with simplicity in describing the system, explaining that its purpose was to comment on the household's activities in a way that might lead to interesting reflections, and answered any questions that the volunteers had. In addition, we described what each sensor measured as we installed the system, and elicited the volunteers' help in positioning them in optimal locations. In this way, we hoped to demystify the system for them, without unduly prejudicing them about its meaning for their lives.

Living With the System

We deployed the Home Health Monitor in a north London household for three months in the early part of 2008. The household consists of a married couple, S and D, and their two cats. Both S and D are in their 30s. S works as an investment analyst in the City of London while D works as a masseuse in a number of health care clinics. Their home is an apartment in a large four-story early 20th century townhouse converted to shared occupation, principally comprising a lounge, a single bedroom, a study, a kitchen and a bathroom. In common with previous work, and in line with the aims of the current research, we studied a single household in depth using, broadly, ethnographic field research methods, to unpack the details of members' engagement with technology rather than sample in a broader but more shallow fashion. In this section, we discuss S and D's experience of the system in terms of the impacts it had upon their home life, in particular looking at how they interpreted the output of the system and how they came to appreciate (critically as it turned out) the system.

Introducing the Home Health Monitor to the Home

In many respects, the sensor units and the printer were not hard to deploy within the fabric of the home. The units could be discretely positioned near where the sensors were operational. The pressure sensors in the living room, for example, were unobtrusively placed under the sofa cushions with the unit on the floor under the sofa arm. The printer was easy to tuck away in a corner of the kitchen.

D and S from time to time took note of the small displays on the sensor units to check that they were picking up on activity in the home in the manner they imagined was intended. In the early stages of deployment, an entire evening spent lying on the sofa watching DVDs went undetected by the sensors. After a while D and S reasoned that the sensors needed to be repositioned depending on whether they were sitting or lying on the sofa.

Some other small adjustments were made by the team to ease the system into D and S's home. At first the printer was set to output at 8am each morning but the noise it made was found objectionable since the kitchen was adjacent to D and S's bedroom. With the output time re-set to 8.30, D and S did not report the system as being disruptive of their everyday activities. D: 'You get used to the system quite easily. It's easy to adapt to. There's nothing much to it.'

This, however, proved double-edged. As the system was discreet in its operation, it did not exhibit notable behaviours which might incite interest or curiosity. For D and S, it only did something once a day and that was a simple affair: the print out of a single card. In contrast to some of our other studies in domestic settings, D and S did not create any special activities or collective 'rituals' surrounding the Home Health Monitor. The breakfast time print out quickly became unremarkable and did not consistently serve as an occasion for talk between D and S.

Interpreting the Output

D and S's engagement with the system quickly came to have a critical flavour. The output from the printer was rarely taken as an object of curiosity in the productive way we had hoped. Often the aphorisms which were printed out in the first part of the deployment were thought to be just wrong. D: 'Last night we got: what do the cats do when you are out? But we were home'. The ambiguous design of the

output was not usually found provocative. D reads out a collection of cards: 'The only thing that overcomes hard luck is hard work… I am not sure whether it is being sarcastic or not... monotony and death, well-ordered life, an evening in... Is it trying to tell us to stay home or that we are home and shouldn't be?' This ambiguity was not found to be engaging by D and S so much as irksome and an indication that the system was not adequately capable of working out what they were doing.

This understanding of the system was also fuelled by a number of perceived inconsistencies in its behaviour. D: 'One day we were told that, although we were home to roost, life doesn't have to be the same, suggesting that maybe it thought we should get out or make changes. Then the next day we got: east-west, home's best.' This change in output occurred even though, to D and S, there was no obvious change in the household's behaviour.

There were occasions when the Home Health Monitor seemed perceptive. After one weekend described by D as particularly lazy, the system printed out: rabbits have their warrens, we have our homes. D: 'We thought that was surprisingly apt'. On another occasion, when D was sick and lying on the sofa all day, the system's 'we are closer to ants than butterflies' captured how she felt. However, D was quick to put this insight into perspective: 'The day before I got 'beware the barrenness of an easy life' so it could just think I am lying around being lazy'. Indeed, the system's haphazard accuracy and aptness suggested to S the possibility that the output might just be random.

From time to time, the levels that the sensor units displayed entered into D and S's reasoning about the system's behaviour. A print out urging them to stay in more often was received after they had spent the day in the bedroom (where there were no sensors). D: 'We though ahh! The units all showed no activity so it must have thought we were out'. In this way, D and S reasoned about the operation of the system to gain a critical understanding of its failings. D and S examined how the system was operating just so as to obtain a settled account of why it made the errors it did. In contrast to several of our previous studies in domestic settings, however, we did not see D and S intrigued by the technical operation of the system as a matter of interest, curiosity or pleasure in it.

The redesign of the system output that took place during deployment did not improve things for D and S. The photographs printed out often had the same irksome ambiguity that the aphorisms did. Some of the activities in the photos were found anomalous. D: 'I don't like ironing so I am not sure what it is saying to me'. The use of personal photos sometimes created a tension between the meaning that D and S had invested in the image and the reasons D and S imagined the system had selected it. Sometimes D and S juxtaposed the pie-chart depiction with the image to understand the system's behaviour. But, as with the occasional inspection of the displays on the sensor units, this was usually done to diagnose an error.

Overall Appreciation

It should be clear from the foregoing that D and S did not find the output of the Home Health Monitor an intriguing object for interpretation in the manner we had anticipated. The cards commonly seemed inappropriate to what had been going on in the house and, when their content was found to be ambiguous, this was annoying rather than provocative. While D and S reasoned about how the system worked, this was not a sign of deepening appreciation and interest, rather they were trying to find out why it was behaving in the way that it did. When system output was found to be accurate, its 'insights' were often mundane or, perhaps, a lucky chance.

D and S were aware that the team had put a lot of effort into making the system. The crafting and finishing of the sensor housings was remarked upon and appreciated as a discreet presence in their home. Ironically, though, this entered into their less than favourable assessment of the system: all this effort for so little consequence. For D and S, there was an abiding mismatch seen between the sophistication of the technical deployment and the thinness of the output. D: 'You would never imagine that it would require this much work to get so little out'.

In previous studies, we have often observed people make comparisons between our systems and other forms of technology that they are acquainted with. In the case of the Home Health Monitor, however, such comparisons further reinforced an unfavourable assessment. As D put it: 'I just don't see how I could benefit from it. I don't see the point of many of these technologies. Other than being a gadget what's the point? I don't like the idea of a system knowing whether you are home or not, unless you were vulnerable and needed some system looking over you'.

While the Home Health Monitor was easy to accommodate within the home, it was easy to ignore. Specific activities did not form around the use and appreciation of the system. The assessment of the system we have documented was swift to form and did not change over time or with different attempts to make the output engaging. The system did not seem to have any new or surprising behaviours emerge with sustained use. For D and S, the system was on the horns of an intractable dilemma as a viable domestic technology: 'We wouldn't want anything more intrusive or engaging and something that is this low level, this slow burn, gives little benefit'.

UNDERSTANDING (A) FAILURE

As the field trial unfolded, we became aware that D and S's experience of our system was rather different to the experience of volunteers who had taken our previous designs into their homes. We have already mentioned these

differences at a few junctures. This caused us to reflect more deeply and explicitly on our past work and what it is that we had found in our prior volunteers' experience that indicated to us that our systems had been relatively successful. In this section, we discuss the failure of the Home Health Monitor from two perspectives. First we describe *how* the system failed, by discussing the symptoms of this failure: features that distinguished this trial from trials that have been deemed successful in the past. Second, we discuss *why* the system failed, by contrasting aspects of the Home Health Monitor's design, and the process we used in pursuing it, with those of more successful systems.

Symptoms of Success and Failure

Here we discuss symptoms of the Home Health Monitor's failure under four themes: engagement, reference, accommodation, and surprise and insight.

Engagement

Perhaps the most fundamental sign of success is that volunteers engage with a design prototype and continue to do so over time. This manifests itself in a variety of ways. Beyond any explicit declaration of liking (which, after all, might be made out of sheer politeness), we take as evidence such things as an enthusiasm about discussing the design and their experience with it; persistence in use and interpretation over time; suggestions for new enhancements that reflect our original design intentions, showing the prototype to friends; disappointment that the field trial must end, and expressions of desire to own the prototype.

Contrast this with D and S's behaviour. Their discussions of the prototype were often marked by a kind of puzzled hesitancy. They didn't discuss it at length, and sometimes seemed uncomfortable lest they offend us. Over time they stopped finding new ways to talk about the system and their experience. If questioned they might conjecture about possible changes someone (not them) might make, but regularly concluded that these changes would not make the system more appealing to them. There was little or no mention of showing the system to friends. Although willing for the field trial to continue indefinitely, it seemed this was because the system was not disruptive rather than because it was a valued addition to their home, and when we did arrange to remove it D and S didn't show any regret.

Reference

A form of engagement that has been striking in earlier field trials involves the tendency for volunteers to discuss successful prototypes through reference to other technologies or experiences that they like. For instance, one of the volunteers who tried an earlier prototype, which displayed aerial photography of England and Wales that moved over time [5], compared it at various times to a hot air balloon, to late night television broadcasts of satellite imagery, and to a plane ride he took in which he spent most of his time in the toilet because it had a window looking down on the earth below. The prototype didn't literally emulate these experiences, nor was it meant to (in fact, we didn't know about them when we designed it). Instead, such references emerged over several conversations, ultimately constituting a category of valued experiences that could include the prototype and thus allow its appeal to be understood and articulated.

When D and S discussed the Home Health Monitor, in contrast, they made relatively few references to related technologies or experiences. When they did, these tended to be ones that they disliked or about which they were suspicious. For example, we discussed surveillance in public areas, with D and S expressing concern about justification, accuracy and potential invasions of privacy, or more private forms of surveillance (such as tracking children's whereabouts) which D and S again found distasteful. On the whole, the Home Health Monitor seemed to be situated with respect to other reference experiences to a lesser degree than we have found other, more successful systems to be, and when it was the references were not flattering.

Accommodation

A notable feature of previous deployments is the degree to which people accommodate successful designs to their existing domestic activities and rhythms. Despite the fact that most of our prototypes are meant to introduce unfamiliar content and interactions to the home, when volunteers persist in engaging with them over time, they tend to find patterns of use that accommodate them within the activities of home. For many of our prototypes this involves periodic engagement during breaks from more purposeful household activities. The status of a prototype might be checked first thing in the morning and then periodically during the day, for example, with more protracted engagement in the evening as an explicitly acknowledged alternative to television viewing. The domestication of a new prototype appears to be a prerequisite to, and evidence for, its success.

From our discussions with D and S, it appeared that they never accommodated the Home Health Monitor in this way. Although they reported looking at its output and interacting with the component sensors regularly over the course of the trial, these activities never seemed to become integrated as part of the household rhythms. It is not that the system disrupted those rhythms. Rather, it seemed to stand outside them, and engagement with it was persistently motivated in terms of participation in a research trial rather than as a form of activity had meaning within the home.

Surprise and Insight

Volunteers persist in using successful systems over time, interpret them with respect to other favoured experiences, and accommodate them to their domestic routines. More than this, successful systems are those which continue to

occasion new surprises and new insights over the course of encounters with them. For instance, new content might appear, or unfamiliar, potentially rare, behaviours might be observed, and this might give rise to new perceptions of the system or the things it indicates. Equally, people may find new meanings for relatively rich but unchanging experiences. Of course, surprise and insight are neither properties of the system per se nor of the people who use it, but instead characterise the relationship between the two. People may perceive novel system behaviours as surprising or not, and such behaviours may occasion new insights or they may not. To the degree that surprise and insight are achieved over the course of a trial, however, a given prototype will tend to be seen as successful.

Although the Home Health Monitor exhibited new behaviours over the course of the trial – not least because we redesigned the output towards its end – D and S did not greet many of these changes as exciting or conducive to new insight. They fairly quickly began to interpret the system's output either as accurate and thus redundant, or as inaccurate and thus uninteresting. After a time, new system outputs did little or nothing to change this situation. Thus a vicious cycle was formed: without new surprises there was little incentive to persist with the system and accommodate it into ongoing routines, and without such accommodation and engagement new insights were unlikely to be pursued.

Why did it go Wrong? Features of Design

As it became clear that the Home Health Monitor deployment was going badly, it became natural to speculate about reasons for its failure. In this section, we discuss potential causes in three sections, corresponding to those we see as relatively specific to the Home Health Monitor, those relating to the particular tactic we used for relating inputs to ouputs, and those having to do with the process we used to develop the system. For the latter sections in particular, we again find that comparing this system with previous ones is a fruitful source of insights.

Features of the Specific Design: No 'Sweet Spot'?

Integral to the design of the Home Health Monitor was the notion that there might be a 'sweet spot' between the extremes of randomness and accuracy in the system's interpretation of household events. If the output of the system fell between these extremes, we speculated, it would correspond well enough to people's perceptions to be meaningful, while remaining different enough to avoid being obvious. Judging from D and S's reactions, however, the Home Health Monitor failed to achieve this. When its output was noticeably linked with household events, it was uninteresting because obvious. Outputs that fell short of this were simply perceived as erroneous. There was little sign that any readings were perceived as lying in between these extremes, or that (partial) inaccuracy might be seen as intriguing.

One of the reasons for this might be that the system failed to establish an independent 'voice' that could be respected even if it disagreed with D and S's perceptions. As with the original Home Health Horoscope, the system's readings usually implied both a diagnosis of the household state ('you are at home more than normal') and an implicit judgement about this state ('home is where the heart is'). In the original deployment, we saw instances in which the volunteers accepted the former, while rejecting the latter, as part of their engagement with the system. For the Home Health Monitor, in contrast, D and S often perceived the judgements as inconsistent. That is, one day the system might imply that staying at home was good, while the next day it would criticise them for not getting out more. Technically, this is because the system did not incorporate a memory for previous household states or output readings, so readings were independent from day to day. The resulting incoherence seemed to undermine D and S's perception of the system, causing them to abandon attempts to make sense of readings that did not agree with their own perceptions.

Of course, there are any number of alternative explanations for why we did not demonstrate the notion of user appropriated interpretation in this deployment. For instance, the first household involved a number of people whose activities were not always known to each other, so that accurate reports might not be redundant, and inaccurate ones difficult to hold to account, and this might tend to make the 'sweet spot' of ambiguity larger than for D and S's household. The horoscopes we used for the first system implied a wider range of contexts and expressed judgements much more ambiguously than the readings that we wrote, making it less likely that anybody would spot clear inconsistencies among them from day to day. Finally, differences in the orientations of the different sets of volunteers may have made them more or less open to engaging with the system playfully rather than analytically. All this highlights the fact that what we assessed was not a system *per se*, but a system deployed in certain context. Thus our failure to demonstrate a 'sweet spot' for interpretation with this system does not disprove the User Appropriated Inference concept. It does however indicate that establishing an intriguing middle ground between randomness and accuracy is not always a simple matter.

General Design Tactic: Widening v. Narrowing Information

Over the course of the trial, D and S periodically remarked about the amount of technology involved in the having the Home Health Monitor in their home. After all, we had installed sensor devices in most of the rooms in their apartment, and these devices were purposely designed to be highly visible, with relatively large form-factors and noticeable, light-emitting displays. What became clear was not just that the technology itself was noticeable, but that it was especially so given that the system output consisted

merely of single sentences emitted once a day. The impression was clearly one of a great deal of equipment being used to very little effect.

Reflecting about this, we realised that this feature distinguished the Home Health Monitor from most of our other designs. Many of the other systems we have developed use relatively simple inputs (distribution of weight, windspeed and direction, information about passing aircraft) as an index controlling the display of much richer sets of information (aerial photography, location-based advertisements, views of the world). The effect is one of *information widening*, in which a small set of locally-determined information gives access to a much richer and geographically wider set of information. This widening, and the simple mapping used to achieve it, appears effective in opening both sets of information to a variety of interpretations.

As we have designed them, the Home Health systems, in contrast, exemplify a form of *information narrowing*, in which relatively rich information gathered from sensors in the home, amounting to hundreds or thousands of readings each day, are progressively averaged, summarised and interpreted to choose a single wellbeing metric – basically, a label – meant to represent the day's activity. The outputs that are eventually used based on this metric are more complex than its label alone, but because that label is the only information used to select them, there is little chance that they will correspond in any meaningful way to the original information. An output sentence or image is essentially a noisy representation of the label or a misleading elaboration on it. From this perspective, the Home Health systems embody a markedly different strategy than most of the systems we have designed.

Our point here is not that information narrowing is always a bad strategy. Indeed, many-to-few mappings are essential to many contributions in ubiquitous computing, e.g., when it is important to ensure robust sensing in the face of individual sensor failure or when several sources of data are required to identify a single feature of interest. Furthermore, systems that can summarise and categorise large amounts of data might even, if user-appropriated interpretation can be achieved, be conceptually provocative. However, recognising the contrast with previous systems does lead us to speculate that finding a 'sweet spot' for systems that reduce information may be far more difficult than encouraging variable interpretation through systems that create lawful links from relevant local activities to bodies of external information.

Design Process: Proving a Point…

As we have discussed, the notion of user-appropriated interpretation was the primary motivation for, and ultimately defined, the Home Health systems. Because of this, demonstrating the concept became the most important factor in proposing and assessing design ideas for the system. This had unfortunate consequences both for our design *process*, and for the *type* of system we ultimately produced.

During the design of the Home Health Monitor, our concern with the experience it might offer (as opposed to, say, its technological implementation or aesthetic presentation) centred almost exclusively on the promise of user-appropriated interpretation. This seemed to distract us from engaging critically with other fundamental questions, such as whether the experience offered by the system would actually be meaningful in a domestic setting. Apart from any theoretical concerns or critical frisson, would there actually be any appeal in living with a system that commented periodically on household activities? The answer, at least for D and S, appears to be 'no' – and we might have guessed this had we asked ourselves the question more often during the development of the system.

This focus on a particular conceptual point, to the detriment of other possible concerns, can be seen as an example of *design for research*. That is, our design activities were pursued primarily in service of a theoretical concern. This contrasts with our more typical stance of *design as research*, in which conceptual payoffs follow from design activities that balance multiple concerns to produce compelling experiences. The result of this was that, paradoxically, although the Home Health Monitor was designed around notions of user appropriation, the design itself did not offer much possibility for alternative forms of engagement. With its strong focus on user-appropriated interpretation, it was essentially a single-issue system.

Attributing failure to designing *for*, rather than *as*, research is not incompatible with the other causes of failure we have discussed, nor to other possible factors such as the particular participants who used the system. On the contrary, the design decisions we made were influenced throughout by this stance towards our design activities. Most generally, it skewed our ability to reflect critically about the Home Health System as a concept, leading us to evaluate the idea as a demonstration of the user appropriated inference concept rather than as a system to be experienced and engaged with in everyday life. From this perspective, we speculate that if we had concentrated more on designing a system that would be compelling to users, we might not have demonstrated the concept so clearly, but we might not have failed, either, and might have learned new lessons from whatever success we did manage to find.

LEARNING FROM OUR MISTAKES

Nobody enjoys failing, yet there was a surprising feeling of relief when we finally admitted to ourselves that the Home Health Monitor was not a success. It allowed us to stop our anxious vigilance for signs of hope and instead to consider what was different about this deployment from previous

ones. In this paper, we have discussed these differences both in terms of symptoms of failure and the features of design that appeared responsible.

To be clear, we are not suggesting that the features we discuss are necessary or sufficient for identifying success or failure. We certainly did not need to articulate them to assess the Home Health Monitor deployment: it was an obvious, incontrovertible and multidimensional flop[1]. Nor do we regard our discussion of engagement, accommodation, reference and the rest as a definitive, complete, finished list, much less a recipe for success. Nonetheless, while consideration of such features should not take the place of the situated judgements made during field trials, they may help to orient attention and support articulation.

The themes of engagement discussed here have arisen from the study of interpretive systems, but we believe they are also relevant for more conventional, utilitarian ones as well. After all, even the most mundane tool will be valued not only to the degree that it solves a problem, but for its ability to evoke enthusiastic engagement, to be understood as congruent with other valued experiences, to fit with ongoing activities, and to suggest surprising new possibilities. Indeed, even the most usable and efficient tool may fail insofar as it is perceived as uninteresting, disruptive, and evocative of undesirable things and experiences. Equally, our speculations regarding the sources of our failure – the inability to establish a 'sweet spot' for evocative output, the use of information narrowing versus widening, and particularly the pursuit of design for research rather than design as research – may also be relevant for narrowly utilitarian systems as well as open-ended ones. If this is the case, then our study of how an open-ended, interpretive system failed would be useful in understanding what it means for any system to succeed.

Perhaps most importantly, the work here demonstrates how systems built to support open-ended interpretation and appropriation can fail. This has been something of a dilemma in the past since, if participants interpreted prototypes differently than expected, this could be taken as evidence for the system's interpretative flexibility. In these circumstances, it sometimes seemed that even failure was a success. Given that such systems are being increasingly discussed in HCI, it seems essential that we be able to distinguish successful examples from unsuccessful ones. What we have shown here is that, despite the many ways they can succeed, failure of interpretive systems is indeed an option. We find this reassuring.

[1] The system was not a *complete* flop, however: designing for sensor legibility seems to have successfully increased participants' awareness of sensor activity. We plan to build on this aspect of the work in future projects.

ACKNOWLEDGEMENTS

This work was sponsored by an Intel grant for Supporting Well-Being and Reflection in the Digital Home and the Equator IRC (EPSRC GR/N15986/01). We are grateful to our participants for their time and generosity.

REFERENCES

1. Bødker, S. When Second Wave HCI meets Third Wave challenges. *NordCHI 2006.* (2006).
2. Boehner, K., DePaula, R., Dourish, P., and Sengers, P. Affect: from information to interaction. *Proc. CC '05*. ACM, (2005), 59-68.
3. Cooper, G. and Bowers, J., Representing the user. In Thomas, P. (ed.), *The social and interactional dimensions of human-computer interfaces*. New York: Cambridge University Press (1995).
4. Gaver, W. (2007). Cultural commentators: Non-native interpretations as resources for polyphonic assessment. International Journal of Human-Computer Studies, *65*, 4 (2007), 292-305.
5. Gaver, W., Bowers, J., Boucher, A., Law, A., and Pennington, S. Electronic furniture for the Curious Home: Assessing ludic designs in the field. International Journal of Human-Computer Interaction, *22,* 1-2 (2007), 119-152.
6. Gaver, W., Sengers, P., Kerridge, T., Kaye, J., and Bowers, J. Enhancing ubiquitous computing with user interpretation: field testing the home health horoscope. *Proc. CHI '0*7 ACM (2007), 537-546.
7. Harrison S, Tatar D, Sengers P. The three paradigms of HCI. *alt.CHI'07*. (2007).
8. Höök, K., Sengers, P., and Andersson, G. Sense and sensibility: evaluation and interactive art. *Proc. CHI '03*. ACM (2003), 241-248.
9. Isbister, K., Höök, K., Sharp, M., and Laaksolahti, J. The sensual evaluation instrument: developing an affective evaluation tool. *Proc. CHI '06.* ACM (2006). 1163-1172.
10. McCarthy J, Wright P: *Technology as Experience*. MIT Press (2004).
11. Romero, M., Pousman, Z., and Mateas, M. Tableau machine: an alien presence in the home. *CHI '06 Extended Abstracts*, ACM, (2006), 1265-1270.
12. Sengers, P. and Gaver, W. Staying Open to interpretation: Engaging multiple meanings in design and evaluation. *Proc. DIS 2006*. ACM Press (2006).
13. Zimmerman, J., Forlizzi, J., and Evenson, S. Research through design as a method for interaction design research in HCI. *Proc. CHI '07.* ACM, (2007), 493-502.