# THE DARMSTADT CHALLENGE
## *The Turing Test Revisited*

J. C. Augusto
*School of Computing and Mathematics at Jordanstown and CSRI, University of Ulster, U.K.*

M. Bohlen
*State University of New York, U.S.A.*

D. Cook
*Washington State University, U.S.A.*

F. Flentge
*TU Darmstadt, Germany*

G. Marreiros, Carlos Ramos
*Polytechnic of Porto, Portugal*

Weijun Qin, Yue Suo
*Tsinghua University, China*

Keywords: Ambient Intelligence, Smart Environments, Turing Test, validation.

Abstract: Significant work has been done in the areas of Pervcomp/Ubicomp/Smart Environments with advances on making proactive systems, but those advances have not made these type of systems accurately proactive. On the other hand a great deal is needed to make systems more sensible/sensitive and trustable (both in terms of reliability and privacy). We put forward the thesis that a more integral and social-aware sort of intelligence is needed to effectively interact, decide and act on behalf of people's interest and that a way to test how effective systems are achieving these desirable behaviour is needed as a consequence. We support our thesis by providing examples on how to measure effectiveness in variety of different environments.

# 1 INTRODUCTION

*"… computers are complex machines that are hard to use. Today we serve them, instead of them serving us. If we are suffering under 1 ton of complexity and inadequacy today, and our machines become 100 times more pervasive in the future we should naturally expect that the complexity and inadequacy of computers will soar 100-fold!…" (Dertouzos, 2001)*

Dertouzos' basic message is that to some extent technology has increased our levels of dependency; we are forced to learn how to use different devices (washing machines, remote controls, computers, answering machines, PDAs, mobile phones, etc.). Whilst machines release us from some tedious tasks and also allow us to do new things, this automation came at the price of introducing other problems which add stress and new complications to humans' lives.

Augusto J., Bohlen M., Cook D., Flentge F., Marreiros G., Ramos C., Qin W. and Suo Y. (2009).
THE DARMSTADT CHALLENGE - The Turing Test Revisited .
In *Proceedings of the International Conference on Agents and Artificial Intelligence*, pages 291-296
DOI: 10.5220/0001665002910296

Ambient Intelligence, "*A digital environment that proactively, but sensibly, assists people in their daily lives*" (Augusto J.C. 2007), promises to change that (Weiser M. 1991, Cook D.J. and Das S. K. 2005, Augusto J. C. 2007, Augusto J.C. 2007). Note that in the definition above: 'Sensible' refers both to accurate diagnosis and timely intervention with emphasis on the users' needs and preferences.

The current challenge then is, 'simply', to satisfy the user. We already have all sort of smart environments exhibiting some degree of intelligence but AmI will not be adopted until the user can use the systems comfortably. Systems should not ask people with Alzheimer's to remember how to use a PDA (or even where it is) or to be dependent on using an accelerometer. Equally undesirable is for these systems to ask people not to carry things when walking over a 'smart floor' so that the system will still know who they are or to rest assured that the video taken in the bathroom will be stored under strict confidentiality in the server.

How can AmI systems be higher quality (e.g., more useful to people)? Here are some open problems the scientific community and companies can focus on to improve things:

- Inferring the emotional/psychological state of a user with high degree of accuracy
- Balancing needs and preferences
- Mediating conflicting preferences in a group

Some work has been done which is more sympathetic with the user's view. For example:

Philips includes a social element as a basic part of their AmI architecture. An exemple of an application which reflects those components of the architecture is the interactive robot called 'iCat' (de Ruyter B. and Aarts E. 2006).

NII-Japan (Richard N. and Yamada S. 2007) has supported research which considers user feedback and preferences to improve acceptability of a reminder system, TAMACOACH: (a) Issues reminders and learns user preferences through User's feedback (accept, later, ignore, done, postpone, cancel), (b) Obtains user's status as an aggregation of: activity level (available, busy , v.busy, away, disconnected); activity context (work, leisure, vacation, commuting, sick, conference, meeting,…); location (office, transportation, home, business trip, …); and mood (v.good, good, average, bad, v.bad,…).

The Polytechnic of Porto (Marreiros G. et al. 2007) is developing an Ubiquitous Group Decision Support System which takes into account the past and current emotions perceived at a meeting. The system is conceived as a multi-agent system with each agent having a perception of other agent's mood and having a role in the algorithm for the negotiation strategy adopted.

MIT's Media Lab (Pentland A. 2005) reported on the use of Hidden Markov Models (HMMs) applied to several contexts including a Smart Car, in collaboration with Nissan-USA. The car monitors the driver's state of alertness and humour which allows the car to adapt its performance to suit the context (e.g., warning the driver about dangers).

In (Streitz et al. 2007) a distinction is emphasized between: (a) System-Oriented, Importunate, Smartness where the system takes/imposes decisions (e.g., 'smart' fridge orders food, sometimes non sensibly), and (b) People-Oriented, Empowering, Smartness where the system makes suggestions (e.g., fridge advises on feasible meals according to fridge content).

MIT (n_house) (Intille S. 2007) proposes to motivate (not to control!) behaviour change by presenting simple messages which are: easy to understand, delivered at an appropriate time and place, using a non-irritating, engaging, and tailored strategy, repeated and consistent.

All these systems (and others we have not listed here) address somehow the issue of providing the user a system which emphasizes technology as a liberating factor for humans and not one that burdens them in a different way. However this efforts are isolated and we feel there should be a common program and agreed goal for the scientific community to work collaboratively in the same direction.

## 2 WHAT CAN BE DONE?

The scientific community has to agree with a standard of measuring user acceptability in terms of intelligent-sensible-sensitive useful metaphors on how we want an AmI system to behave: should an ideal AmI system be more like an ideal butler or like an ideal nurse or like an ideal personal assistant or like …? Lets take one of these metaphors as an exercise. What we require from a human that we think is an 'Ideal' Personal Assistant? Here there is a partial list: is always ready, knows our preferences, knows our needs, is kind, (also entertaining?), knows when to interrupt (…and when not to!), knows where things are (or may be) located, knows how the outside world works (at least on some specific areas like train time-tables, booking tickets, buying food online, etcetera).

Some challenges of course are: (a) how to set up/update this knowledge and its hierarchy of importance?, (b) balancing 'needs' versus 'preferences' (and their change over time), (c) mediating conflicting preferences, e.g., selecting a T.V. channel for a family to watch (sounds familiar?), and (d) how the system can better understand the state of mind of the user(s), e.g., mood.

Still the existence of obstacles does not mean that they are unsolvable and on the other hand many benefits can be achieved by setting up an agenda to test the achievement of such systems. We propose here the setting up of a benchmarking challenge which we call 'The Darmstadt Challenge' (because it was first discussed during (WHAAmI 2007) hosted in Darmstadt). This challenge can be used to measure progress (and eventually achievement) of an AmI system with the desired characteristics.

Part of the definition of the challenge involves measuring user satisfaction and agreement that the system acts as an ideal personal assistant through a questionnaire. A given system will have passed the Darmstadt Challenge when a group of no less than, say, 10 users rank the service as acceptable and human-comparable in more than 80% of the elements assessed.

Given that the challenge has the main ojective of judging to which extent the Smart Environment system exhibits Ambient Intelligence there is a relation with a well known concept in Artificial Intelligence, the Turing Test (see Turing A. 1950, Russell S. and Norvig P., 2003, TT 2007). In the Turing Test, also called sometimes the Imitation Game, an interrogator posts queries that are answered from two different sources. One is a machine and the other one a human. The interrogator should guess which is which. If the machine manages to lead the interrogator to think that the machine was a human then the machine would have passed the Turing Test.

A fundamental difference in between the Darmstadt Challenge and the Turing Test is that we do not measure general intelligence of a system but acceptability on behalf of humans that the system is capable to perform satisfactorily a specific task which requires intelligence.

Although Turing was aware of the fact that the assessment of 'intelligence' includes intuitive elements on the part of the observer, we believe that the inclusion of a 'social' element further distinguishes the Darmstadt Challenge from the Turing Test. The catch word 'social' here should be interpreted beyond the caricature of social (faux social: smiley faces) and towards something reminiscent of meaningful exchange. The premise that a smart system will also be socially satisfying is not necessarily true. Smartness for social settings includes understanding and appreciating limitations, e.g., do nothing where appropriate. Here we assume that 'appropriate' can be clearly identified. Notice for example that 'appropriate' is related to cultural values.

Which elements to monitor? According to (Treur J. 2007) the main aspects of human life to consider are: social, emotional, cognitive, physiological and neurological (see Figure 1). We should add they can be monitored in all the possible combinations and with many different priority systems according to the contexts of applications and the people involved. These aspects of human life are not isolated but rather inter-dependent, which makes their monitoring and understanding to be tasks of formidable complexity.

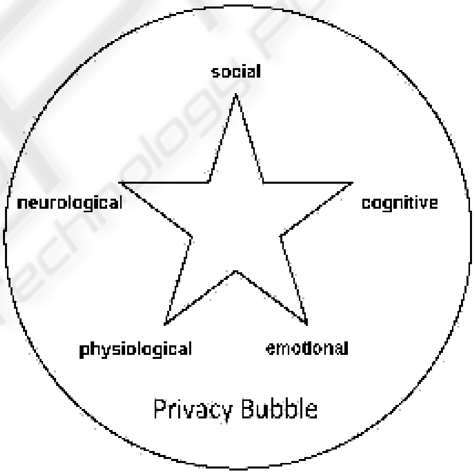

Figure 1: Key aspects of human life and their complex interdependencies.

# 3 EXAMPLES

## 3.1 Smart Home Scenario

A smart home is a home environment that uses sensors, controllers, and software algorithms to acquire and apply knowledge about its residents and their physical surroundings in order to adapt to the residents and improve their experience in the environment (Youngblood G.M. and Cook D. J. 2007). What, then, is a good analogy that we can use to convey the desired features of a smart home?

Here we propose the metaphor of a smart home as a silent, ever-present, valued butler.

What comes to mind when we think of a home's butler? The persona of a butler is depicted as being discreet and unobtrusive. A butler has lived with the family for years and so is sensitive to the master's whims, needs, abilities, and habits. Instead of responding only when called, a good butler is always available and anticipates his master's requests. He does not attempt to perform every task and solve every problem for his master, but over time learns the types of tasks that are needed and how best to perform them. The people who live in a smart home will be more comfortable and more productive because of the presence of their butler.

See a sample of evaluation form in Appendix A.

### 3.2 Smart Public Spaces Scenario

Independent of the issue of a metaphor one would have to distinguish between the design of such a system and the experience of it. Two separate problems. We can see the technical system as a foreign presence, an 'other', one that can work for us (also in ways we do not expect) and that it is difficult to understand at times (which is often the case). This has to be coupled with well designed transparency so that we understand intuitively how to interact with it to make the system effective.

An example of such systems in the public space could be something like a 'cyborg taxi driver' capable to engage in an interesting dialogue. In general the metaphor should suggest itself through the system's abilities and, to some, an abstract (even vague) notion of an 'other', can be satisfying.

See a sample of evaluation form in Appendix B.

### 3.3 Smart Office Scenario

A metaphor that is more appropriate to a smart decision room is an ideal secretary, 'someone' that is there to assist the user or even to act on the user's behalf. At a smart decision room we have two distinct levels of assistance: software and infrastructure. The software available at the smart decision room should assist the different participants in the decision process. For example, the ideal secretary should be able to suggest pertinent arguments, to advance the trends of the meeting alternatives and to analyse if the preferred or undesired user alternatives have possibility to win (or not), or if the user is unavailable. The ideal secretary may have some autonomy and take some actions on the participants' behalf. At the infrastructure level there are a set of devices that contribute to the creation of a smart environment (e.g. the ideal secretary should turn off the lights of the room when someone is making a presentation)

See a sample of evaluation form in Appendix C.

### 3.4 Smart Classroom Scenario

A Smart Classroom is the environment that manages both the classroom and the interaction and motivation elements to support the delivery of a lecture or other pedagogical material. Here the metaphor can take the form of a teaching assistant which is full time dedicated to support teaching and learning and can make observations and take decisions in real-time to achieve that goal. Typical Smart Classrooms include enhanced interaction between students and teachers through tablet PCs and instant connectivity which allows sharing of, often anonymized, answers, statistics on students perception on a topic, preferences etcetera. Exemplar cases (see for example, Shi Y., Xie W., Xu G., Shi R., Chen E., Mao Y., and Liu F., 2003) include also intelligent systems which can use automatic focusing of video on teachers and/or students actions as well as intelligent use of voice processing to facilitate the use of the available technology on behalf of the user.

See a sample of evaluation form in Appendix D.

## 4 A GENERIC ASSESSMENT METHODOLOGY

Out of the isolated and specialized assessment methods for different environments listed above we can distil a general methodology that can be applied to different environments:

1) Define a set of characteristics $\{c_1,...,c_n\}$ which are expected/shown by AmI systems. Let us assume for the time being these characteristics can be extracted through a questionnaire and quantified.

2) Each situation $S$ demands specific profiles of these characteristics $\{c_1=x_1,\ ...,\ c_n=x_n\}$ (maybe several profiles are possible, maybe also user-specific profiles, etc). These profiles have to be defined by the potential users. This could be done using questionnaires and an appropriate operationalisation of the characteristics

3) The user interacts with the system and measure the characteristics expressed by the system, again using questionnaires (this may also require to assess the importance of each characteristic).

## 5 CONCLUSIONS

Significant effort has been devoted to the advancement of systems related to Pervasive and Ubiquitous computing and Smart Environments. The advances so far have not made these systems accurately proactive. On the other hand a great deal is needed to make systems more sensible/sensitive and trustable (both in terms of reliability and privacy). For example, the Robot@Home challenge set up as part of the RoboCup competition is mainly focused on the skills of a robot to navigate a house.

We put forward the thesis that a more integral and social-aware sort of intelligence is needed to effectively interact, decide and act on behalf of people's interest. We proposed a specific challenge devoted to measuring how close an AmI system is to the ideal system for a specific user. Although we generally compared that with matching the ideal of a personal assistant we exemplified how in different scenarios more specific metaphors can apply. We called this process of looking for an evaluation framework and its application: *the Darmstadt Challenge.* Although it may bring resemblance to *the Turing Test* it is a different mechanism with a different goal. The goal of the proposed challenge is much more utilitarian than the general goal of the Turing Test. The process is also different in the sense that the users know where the computer is and what is trying to achieve.

We sustain that even if the test is not perfect there will be substantial benefits from exercising the test as the benchmark in the field. We hope this will stimulate a discussion within the community to both, further refine the Darmstadt Challenge and to make it systematic. Its sustained application hopefully will contribute to the improvement of systems and ultimately to achieve the aim that artificial systems truly serve humans and not vice-versa.

## REFERENCES

Cook D. J. and Das S. K., 2005. *Smart Environments: Technology, Protocols and Applications*. Wiley-Interscience, 2005.

Augusto, J. C., 2007. Ambient Intelligence: the Confluence of Ubiquitous/Pervasive Computing and Artificial Intelligence. In *Intelligent Computing Everywhere*, pages 213–234. Springer Verlag, 2007.

Augusto J. C. and Cook D. J., 2007. *Ambient intelligence: applications in society and opportunities for AI (tutorial lecture notes)*. IJCAI-07.

Augusto J. C. and McCullagh P., 2007. Ambient Intelligence: Concepts and Applications. Invited Paper by the *Int. Journal on Computer Science and Information Systems*, V 4, N 1, pp. 1-28, June 2007.

Dertouzos M., 2001. Human-centered Systems, in *The Invisible Future*, Denning (Ed.), pp. 181-192.

Intille S., 2007. 'Smart People, Not Smart Homes', *Proceedings of ICOST 2006*, Nugent and Augusto (Eds.), pp. 3-5. IOS Press.

Diener E., Emmons R.A., Larsen R.J., and Griffin S., 1985. The Satisfaction with Life Scale. *Journal of Personality Assessment*, 49, 71-75, 1985.

Marreiros G., Santos R., Ramos C., Neves J., Novais P., Machado J., and Bulas-Cruz J., 2007. Ambient Intelligence in Emotion Based Ubiquitous Decision Making. *Proceedings of the 2nd Workshop on Artificial Intelligence Techniques for Ambient Intelligence (AITAmI'07)*, Augusto and Shapiro (Eds.), pp. 86-91, Hyderabad, India, 2007.

Pentland A., 2005. Perceptual Environments, in *Smart Environments*, Das and Cook (Eds), Ch. 15. Wiley. 2005.

Richard N. and Yamada S.,, 2007] Context-Awareness and User Feedback for an Adaptive Reminding System. *Proceedings of the 2nd Workshop on Artificial Intelligence Techniques for Ambient Intelligence (AITAmI'07)*, Augusto and Shapiro (Eds.), pp. 57-61, Hyderabad, India, 2007.

RH, 2008. http://www.ai.rug.nl/robocupathome/

Russell S. and Norvig P., 2003. Artificial Intelligence: A Modern Approach (2nd Edition). Prentice Hall.

de Ruyter B. and Aarts E., 2006. Social Interactions in Ambient Intelligent Environments. *Proceedings of 1st Workshop on Artificial Intelligence Techniques for Ambient Intelligence (AITAmI'06)*, Augusto and Shapiro (Eds.), pp. 3-4. Riva del Garda, Italy.

Shi Y., Xie W., Xu G., Shi R., Chen E., Mao Y., and Liu F., 2003. The smart classroom: Merging technologies for seamless tele-education. *IEEE Pervasive Computing*, v2.

Streitz N. et al., 2007. Smart Artefacts as Affordances for Awareness in Distributed Teams. N. Streitz, T. Prante, C. Röcker, D. van Alphen, R. Stenzel, C. Magerkurth, S. Lahlou, V. Nosulenko, F. Jegou, F. Sonder, D. A. Plewe. In *The Disappearing Computer*, pp. 3-29. Streitz et al. (Eds.). Springer Verlag. 2007.

Treur J. 2007. Introduction to Proceedings of (WHAAmI 2007).

TT 2008. http://plato.stanford.edu/entries/turing-test/

Turing A., 1950. Computing machinery and intelligence, Mind LIX(236): 433-460.

Weiser M., 1991. The computer for the 21st century. M. Weiser. *Scientific American*, 265(3):94–104, 1991.

WHAAmI 2007. *Proceedings of the First Int. Workshop on Human Aspects in Ambient Intelligence*. Tibor Bosse, Cristiano Castelfranchi, Mark Neerincx, Fariba Sadri, Jan Treur (eds). Darmstadt, Germany, 2007.

Youngblood G.M. and Cook D. J., 2007. Data mining for hierarchical model creation. G.M. Youngblood and D.J. Cook. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 37(4):561-572, 2007.

## APPENDIX A

Sample of evaluation for a Smart Home:

- Does the introduction of AmI technology change the look or feel of the house?
- What changes in daily life are needed to make use of AmI technology?
- For how much of the house is smart home assistance available?
- How much effort is required to request assistance from the home?
- Does the quality of the assistance increase with use and time?
- Does the assistance customize itself to the residents of the home?
- Does the assistance improve your productivity at home?
- Does the assistance improve your health and/or safety at home?
- Which aspects of the Smart Home were useful?
- Which aspects were disappointing?
- Would you recommend use of the Smart Home to a friend or family member?

## APPENDIX B

Sample of evaluation for a Public Library:

- Where you able to get what you wanted?
- Did you notice the AmI system?
- If so, how often did you forget that you were in an AmI environment?
- Did the system enhance your visit (to the library)?
- If so, in which way?
- Where you surprised by the AmI system?
- If so, in which ways?
- Do you feel that the AmI system improved the services offered at the library?
- Do you feel that the AmI system made the public space a better space?
- If so, in which way?
- Which aspects of the AmI system you experienced were disappointing?
- Which aspects of the AmI system you experienced were annoying?
- Would you return to this library because of the AmI system you experienced?
- Would you recommend the library to a friend because of this AmI system?

## APPENDIX C

Sample of evaluation for a Smart Office:

- Are the devices existent at the smart decision room appropriate to the group decision process?
- Did the interactions with the 'personal assistant' introduce some 'noise' in the decision process?
- Are the autonomous interactions of the 'personal assistant' with the environment synchronized with the meeting status?
- Has the personal assistant an ethical behaviour?
- Does the use of a personal assistant improve and facilitate the uses of Group Decision Support System?
- In general how do you classify the interaction with the personal assistant?
- The argumentation structure and strategy suggested by the personal assistant is pertinent?
- How do you classify the information that the personal assistant collect about the other meeting partners?
- How do you classify the introduction of emotional processes in the personal assistant design? Are the emotional aspects relevant in the decision process?
- How do you classify the behaviour of the personal assistant, proper or too invasive?

## APPENDIX D

Assessment for a classroom assistant:

- Always ready to help both, teacher and students.
- Successfully coordinates all the devices, classroom devices or personal devices inside the classroom.
- Understands the teacher's multi-modality commands: gesture, voice or laser pen, etc.
- Enable the student outside classroom to communicate with local students.
- Manage an electronic whiteboard for sharing notes with local and remote students.
- Knows the teacher's preference: class-schedule, habits.
- Motivate the students to communicate with each other, and with the teacher.
- Allow the students to ask questions without necessarily interrupting the teacher.
- Record the classroom for later reviewing.