

The hidden dangers of experimenting in distributed AI

Annika Smit *

Masja Kempen

Niek Wijngaards

DECIS Lab

Thales Research & Technology Netherlands

P.O. Box 90, 2600 AB Delft, The Netherlands

+31 15 2517{865, 861, 867}

[annika.smit, masja.kempen,
niek.wijngaards}@icis.decis.nl](mailto:{annika.smit, masja.kempen, niek.wijngaards}@icis.decis.nl)

André Meyer

DECIS Lab

TNO Defence, Security & Safety

P.O. Box 96864

2509 JG The Hague,

The Netherlands

+31 70 374 01 78

andre.meyer@{decis.nl, tno.nl}

ABSTRACT

Research on multi-agent systems often involves experiments, also in situations where humans interact with agents. Consequently, the field of experimental (human) sciences becomes more and more relevant. This paper clarifies how things can and often *do* go wrong in distributed AI experiments. We show the flaws in methodological design in existing papers (both with and without humans) and work out an example involving human test-subjects to introduce the fundamental issues of experimental design. Furthermore, we provide researchers with an approach to improve their experimental design. We wish to stimulate researchers to conduct better experiments – which will benefit us all.

Categories and Subject Descriptors

C.4 [Performance of Systems]: measurement techniques

General Terms

Measurement, Performance, Design, Experimentation, Human Factors, Standardization, Theory.

Keywords

Human-agent experimental design.

1. INTRODUCTION

In general, computer systems are built to perform on behalf of human users. Even autonomous software agents depend on human interaction for critical actions, because they are not (yet) seen as legal entities. Validation of human-agent interaction through experimentation is getting more attention lately in AI and computer science. Research in multi-agent systems more often involves humans as full-fledged participants in mixed-initiative interaction, for example in negotiation (e.g. where humans can set orders), decision support, workflow systems, etc. Actual empirical (human) research is conducted – which is an interesting and valuable trend – but needs to produce useful results.

Papers presenting experiments are seen more frequently as beneficial to our whole research community. At the same time, this community does not have a tradition in experimental design,

especially not with humans. Valid and objective results are not easily gathered. Experiments need to be set up in such a way that peculiarities in human nature do not interfere with the experiment at hand. Psychology as a discipline has invested much time and effort into human experimenting – sound experimental methodology thus is available. It would be wise to apply this knowledge to experimentation in distributed AI.

Empirical research on human subjects is time consuming. It involves going through several necessary steps; Explicitation of the research question, or *hypothesis* formulation. Choice of testing method, or design. Decisions about *how* and *what type* of data will be gathered. What are the units of observation, and how many of them? Choice of data analysis methods. These issues are all crucial for choosing an *experimental design*, which refers to the formal structure of the experiment. The design *must* have at least the following properties: a) comparison of the observations under different levels of the independent variable should lead to non-ambiguous conclusions, and b) the comparison should be significantly testable.

Due to restrictions on the length of this paper we omitted the section on a brief overview of good experimental design. Instead we refer to [1], wikipedia.org, or any other good book on experimental design and statistics. Also unknown terms mentioned without explanation can be considered to be easily looked-up in such experimental handbooks. The important lesson to take home is that designing good experiments, best suited to your research question, is an art in itself. Invest time in it or you might end up with a meaningless experiment and a lot of time wasted. A final point we would like to make is that experiments should be reported in such a way that others can replicate the study.

In this paper, a number of erroneous experimental designs found in the latest AAMAS 2005 proceedings are used to illustrate common experimentation errors. We do not cite these researchers in detail, but appropriately name them ‘Study 1, 2, etc’. Our aim is merely to use them as examples of well-intended, but wrongly applied efforts. Using a fictitious case, we show how proper experimentation methodology has to be applied and explain how to avoid some common pitfalls.

2. ERRORS in HUMAN EXPERIMENTS

We now focus on flawed human experiments to point out possible improvements. We confined ourselves to the papers of AAMAS

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

AAMAS’06, May 8-12, 2006, Hakodate, Hokkaido, Japan.

Copyright 2006 ACM 1-59593-303-4/06/0005...\$5.00.

* Annika Smit currently works at ISC, The Netherlands;
annika.smit@isc.politie.nl

2005, to show the relevancy of this subject. As we do not want to discredit researchers who are doing interesting research, we anonymously refer to the studies, which are numbered randomly.

People are adaptive systems who are constantly reacting to their environment and thus are prone to all kinds of influences – also the ones we do not want to measure. If a comparison is made between conditions (e.g. different interfaces) in terms of quality (e.g. people can work better with a certain kind of display), the different conditions cannot be presented in one specific order to all subjects. People are known to react differently in the beginning of an experiment than near the end. Numerous different factors play a role, such as practice effects, tiring effects, enthusiasm, time-of-day effects, etc. Therefore, the absolute basics of experimental design include a proper ordering of conditions. Unfortunately though, there are experimental studies that lack even this essential basic property [10]. Sometimes there seems to be a design, but it is not made explicit enough [6, 9].

Subjects need to be *randomly* assigned to the conditions. People differ from each other in many ways; these differences need to be averaged out by appointing people to different conditions in a random manner. Also important is what psychologists call ‘external validity’: The type of subjects used determines whether the findings can be extrapolated to other populations. It is a common mistake to use colleagues for experimentation [8]. Also the number of subjects is important. It is not useful to compare different conditions while using only two or three subjects. This may seem obvious, but experimental populations of this size are not uncommon at all [10].

Obviously, the period before actually conducting research entails the important basics of testing. But things can also go wrong in later stages. The description of procedures in papers is often vague and lacks important details. Methods and procedures are often not orderly described. However, standardization of method formats is common in many research areas. A clear format is the following: *Participants, Design & procedure, Material and stimulus presentation, Results (Statistical analysis), Discussion.*

Statistics are related to the aspect of subject amount. If only two or three subjects are tested, results can never be statistically significant. Unfortunately, in distributed AI research the specific type of statistical analysis used is often not stated [9]; sometimes no statistical analysis is done at all [10, 8].

3. STUDYING HUMANS PROPERLY

We here give an example. The data reported are fictitious, but realistic (see e.g. [3]). Let’s assume that, for a Dutch auction (in which the price of the goods decreases over time), recommending trade agents have been implemented. These trade agents support the human buyer by indicating when it is the best time to buy. Legislation requires the human still to ‘push the button’, as indication of an actual order to buy; software agents do not have the (legal) authority to make a contract. However, supporting agents can be used which may indicate when to buy.

Whenever the time is ready to buy, the user should be able to push the button *as fast as possible*. The researchers want to know with what type of interfacing human use the trading agents best. The clock used for the auction is shown on the user’s screen. This graphical representation of the clock changes in some way, thereby showing the users the best moment to buy and push a button. An analogous clock and a digital clock are used; both count down prices. Four different conditions are compared: 1)

analogous clock that changes background colour as indication to buy, 2) analogous clock that changes the clock-hand colour, 3) digital clock that changes background colour, 4) digital clock that changes the digits colours. The question is: Is there a significant difference in response time in the different conditions?

Experiment 1

Method

Participants. Eight students/... (female/male, average age) participated in this study (give background information).

Design & procedure. There were four 30-min conditions; a 10-min break followed after each condition. See Table 1 for the order of conditions. Using this type of counterbalancing, four subjects would have been enough, but probably *not* for significant results.

Table 1. Order of conditions (Balanced Latin Square design)

Subject	Order of conditions			
1, 5	1	2	4	3
2, 6	2	3	1	4
3, 7	3	4	2	1
4, 8	4	1	3	2

Material and stimulus presentation. This section should entail a description of all the materials and how stimuli were presented.

Results

A repeated measures ANOVA (Analysis of Variance) was used, with within-subject factor: Condition (the four different clock types). Usually data of one or more subjects cannot be analyzed: Equipment may fail or subjects do not finish the task. This is usually reported in order to inform researchers how many subjects may be ‘lost’ along the way. We assume that data of one of the subjects was ‘lost’. Averaged RT data are shown in Table 2.

Table 2. Mean Reaction Times (RTs, in ms) of seven subjects to the buy-signal in the four different conditions

Subject	Condition			
	1	2	3	4
1	410	460	490	500
2	490	440	450	520
3	430	480	430	480
4	470	410	510	490
5	440	510	470	470
6	460	500	520	440
7	450	420	420	460

It is not enough to just present average results of the different conditions. Let’s see what the data look like when averaged over all subjects (Figure 1). Pay attention to the distorting effect on the y-axis (that starts at 435).

Condition 1 seems to yield the fastest responses, but the ANOVA reveals no significant Condition effect ($p > 0.1$): The independent variable Condition (with four levels in this case) did not have a significant effect on RT (dependent variable).

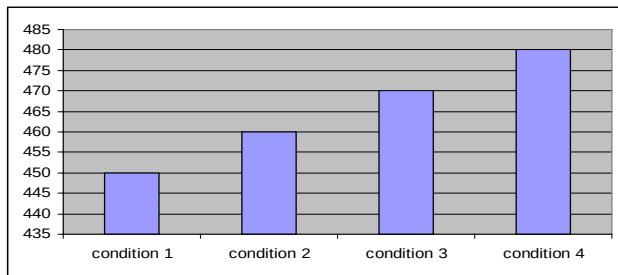


Figure 1. Performance measured in Reaction Times (ms), comparing four conditions, averaged over all seven subjects

Experiment 2: using the same example without methodology

If the four conditions were in the same order to all subjects (condition 1, 2, 3, 4.), the results would have been contaminated by fatigue effects, which have been found in the literature [4]. RTs may slow down 23% for the second and third presentation (compared to the first), and by 49% for the fourth presentation (compared to the first). If we now 'control' the data in Table 2 for these effects, the picture is totally different (see Figure 2).

The ANOVA yields a significant effect of condition; Condition 4 has the fastest response and Condition 1 the slowest [all p-values ≤ 0.01]. So now there indeed is a significant condition effect, but the effect is completely different from the picture in Figure 1. Of course, we do not know whether fatiguing effects occurred. Furthermore, there are a lot more effects that play a role [2]. All these effects are compensated for by using a proper design (see section 2).

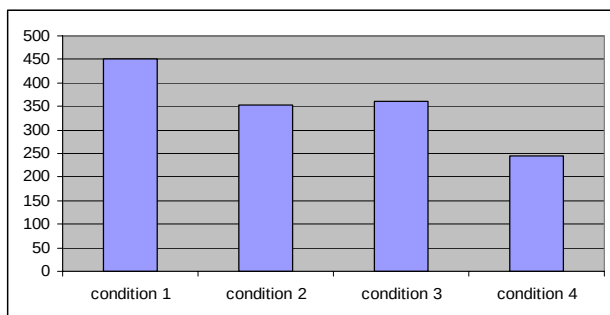


Figure 2. Performance measured in Reaction Times (ms), comparing four conditions, averaged over seven subjects (hypothetically 'controlled' for effect of fatigue, see text)

4. DISCUSSION & RECOMMENDATIONS

Experimentation is extremely useful, as it aids us to better understand strengths and weaknesses of our models, theories and implementations. However, conducting experiments is very difficult. The difference between solid and weak methodology is the difference between valuable data and the absolute waste of time. Statistics can never make up for experimental design flaws. It is beyond the scope of this paper to provide a complete tutorial on sound experimental design. We have provided an introduction, including examples of common errors and a (fictional) case with a better experimental design. To support researchers in improving their experiments, we present a list of relevant issues and actions:

1. Think carefully about what questions you want to answer.

2. Understand that the area of experimental methodology is a specialized field; conducting experiments is very hard.
3. Accept that there are more effects out there that may harmfully influence your experimental results, therefore:
4. *Consult an expert* on methods & statistics (e.g. in the Psychology dept.). Do not try to do it all yourself.
5. Read appropriate literature on the basics of experimenting (ask the expert what would be useful in your case).
6. Let your expert analyze and comment on the design, methodology and statistical analysis you've chosen.
7. Report not only your results, but also your experimental design and procedure and chosen statistical analyses.

It is important that the distributed AI community improves its use of experimental designs. There is already research on software agents that have human-like capabilities, e.g. agents with memories, learning and adaptivity. Although agents do not get 'tired' as humans, they do live in resource-restricted environments (e.g., processing time, memory, etc.), which influence their performance. Conducting experiments with individual or groups of software agents is not an easy job, and many techniques used for human experimentation will also be necessary for experimentation with agents. This overlap in research areas is a topic we wish to explore in future publications and experiments.

5. ACKNOWLEDGMENTS

The research reported here is part of the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, grant nr: BSIK03024 (<http://www.icis.decis.nl>). The authors wish to thank the participants of the 'Creating Bonds with Humanoids'-workshop at the AAMAS 2005 conference, for raising awareness on the important issue of sound experimental design.

6. REFERENCES

Please note that the AAMAS 2005 papers are referred to anonymously; numbers are used in a random fashion (i.e. the order and numbers are not related to the papers they refer to).

- [1] Elmes, D. G., Kantowitz, B. H., Roediger, III, H. L. *Research methods in psychology*, West Publishing Co: Minnesota (USA), 1985, 2nd edition.
- [2] Kraepelin, E. *Die Arbeitscurve* [The working curve]. Leipzig: Wilhelm Englemann, 1902.
- [3] Smit, A. S., Eling, P. A. T. M., & Coenen, A. M. L. Mental effort causes vigilance decrease due to resource depletion. *Acta Psychologica*, 2004, 115, 35-42.
- [4] Smit, A. S. *Vigilance or availability of processing resources: A study on cognitive energetics*. Dissertation Thesis, Nijmegen, 2004, Chapter 3.2, pp 39-54.
- [5] Study 1, in *Proceedings of the AAMAS 2005*, by ACM Press.
- [6] Study 2, in *Proceedings of the AAMAS 2005*, by ACM Press.
- [7] Study 3, in *Proceedings of the AAMAS 2005*, by ACM Press.
- [8] Study 4, in *Proceedings of the AAMAS 2005*, by ACM Press.
- [9] Study 5, in *Proceedings of the AAMAS 2005*, by ACM Press.
- [10] Study 6, in *Proceedings of the AAMAS 2005*, by ACM Press.
- [11] Study 7, in *Proceedings of the AAMAS 2005*, by ACM Press.