Reviews of the Autonomy papers

in preparation for the written part of the contextual supporting area of the qualifying exams

by Stefan Marti

Version of August 31, 2001, 9:46am.

Note: This is my scratch area; my summaries of the papers, as well as personal comments. These notes are updated very frequently, not only at the end where new reviews get added, but also throughout the whole file.

Sociological/Psychological Issues

Donald Norman (1994). How Might People Interact with Agents

Summary

- Agents: produce *fear*, *fiction*, and *extravagant claims*. However, main difficulties are not technical, but social.
- Agents are *not new*, they descend from *automata* (e.g., autopilots), but they are more human (language, animated graphical appearance). Seem to have goals and intentions ((Huh?? An autopilot might have that too, it's just a question of complexity?))

Main issues

- How do people **feel** about agents?
- Are people **comfortable** with, and can they **accept** their automatic autonomous actions?

- People have to feel in control
- Nature of human-agent interaction
- Safeguards, against runaway computation
- Accurate expectations, no false hopes
- Privacy
- Hiding complexity and being transparent at the same time.

Feeling of control

People need that for their activities. They also need to understand what is going on, and have confidence in the system. That comes slowly, and failing automation inhibits building up confidence.

Norman repeats: in order to accept technology, especially agents, everything "has to go according to plan", both in a **social** way and in a **technical** sense.

Social way: develop a conceptual model of the actions, so that they are understood:

- What actions have been taken on behalf of the user?
- Private matters remain private
- Expensive/unwanted actions are not taken without explicit permission
- Trace all events in past
- Undo unwanted events in past

Comfort level and **trust** may increase over time and with reliable functioning. Agents first have to be transparent so that everybody can see what they are doing, and later they can hide that complete reporting (because it gets annoying) ((This is trivial, because it is true with any new technology, be it a car motor or air-conditioning system.)) But anytime, it has to be possible to go back to the complete reporting level (like with bank records).

Overblown expectations

The more we **anthropomorphize** agents, the more likely we create **false hopes**. Speech reco creates expectations of language understanding, having goals creates expectations of understanding human goals. Norman thinks there are no moral problems as long as there are no false promises, no deception (which is controversial). People need a "system image" to avoid that.

Safety

This is part of feeling in control. It gets problematic when humans program agents maliciously (viruses).

Privacy

This is also part of sense of control, but probably bigger. Is often about interests of one group of people versus another. Agents might have personal information of the owner, and they have to act responsibly with this information. Again, it's a question of trust.

Human-agent interaction

What is appropriate form of interaction between agent and person? How to instruct, what feedback, how to get the conceptual model of the agent's actions, how the agent gives advice to person?

Concluding remarks

A once unleashed technology will not go away, so agents will stay ((see Joy)). Agents simplify interaction with technology (computer, car), provide friendly assistance (so smoothly that humans are not even aware of it, like automatic choke in cars, etc) But agents are more powerful than other artifacts, since they can have some level of intelligence, some form of self-initiated, self-determined goal ((Not sure, depends on the definition. I think that's just increased complexity of the system))

Comments

There are some technical issues: agents have no more intentions and goals than an autopilot might have (at least, an idealized autopilot)--it is just a question of complexity. Or can autopilots NOT be agents?

His suggestion that people have to get used to agents, and that it takes time to trust them, is also trivial, since developers have to do that with ANY new technology: they first monitor very closely what the new machine is doing, and when they trust it, they want to see only higher level reporting. That is not new with agents. The question is if users trust the developers the same way car drivers trust the car manufacturers that the car doesn't explode suddenly...

His views of what agents can do and will do are way too limited. One could actually replace the term "agents" with "slaves", and it still works what he says. But autonomous systems are not always here to do work for us in such a limited sense. What if they are here to entertain us, to love us, or to fulfill other social functions? It seems to me that these issues are not covered in Norman's view.

Another class of problems will appear when agents start to claim to be a non-carbon based life form.

There are other shortcomings. In terms of user interface, we already have experience with autonomous entities--this is not new to us. Not only are all humans around us autonomous entities, but also all animals (especially pets). There is no need to find a new "programming model", as long as we treat the autonomous entities similarly to people and/or animals. (Although there might be certain situations where other models might be more efficient though, but that is only true for less intelligent agents.)

He complains that agents will have "some level of intelligence, some form of self-initiated self-determined goals." He is not explicit, but he almost makes it sound like these characteristics would be bugs. Does he have a problem with the billions of humans and other life forms on earth that have "some level of intelligence, some form of self-initiated self-determined goals"? I doubt it. So why can't we accept highly intelligent agents as artificial life forms that have these capabilities?

I think Norman's stereotypical human must be a "control freak." That might be a very significant bias, given that not everybody is a control freak.

Jonathan Steuer (1995). Self vs. Other; Agent vs. Character; Anthropomorphism vs. Ethopoeia

Summary

Purpose of this study: **compare responses to human with responses to human-like agency**.

Structure of this paper: four sections:

(1) Typical reactions to humans. From Social Psychology, Sociology.
(2) How does A.I. look at concept of agent and agency? From A.I.
(3) Anthropomorphic responses to technology. From HCI.
(4) Can we use theories of conversation to design agents? From Psycholinguistics.

1. Typical reactions to humans. From Psychology and Sociology

Idea: *Let's use traditional psychological experiments, replace one human actor with an agent, and then see how humans react.*

Prerequisite: *Source (sender) perception and message perception are linked*. A person's perception of a source influences the person's response to the message, and the other way round. A person's reaction to a message influences this person's perception of the sender of the message!

Proof 1: Self-referential messages ("I am great") are regarded as less accurate and objective than other-referential messages ("He is great"), because my own assessment must be tainted and I might have suspect motives. (Self-praise is dubious.)

Proof 2: An identical message ("That was excellent performance") directed to somebody else is perceived as friendly, but directed towards oneself as unfriendly.

This distinction between self vs. other might also apply for interaction between human and technology.

Technology as autonomous sources

Literature seems to agree on that normal adults do not apply social rules to technologies: technology can't be an autonomous source for messages. Why does it still happen?

Explanation 1: If adults still do, they are abnormal, exhibiting para-social interactions, and suffer from socio-emotional problems.

Explanation 2: Adults do not show social reactions to technology, but to the **creator** of the material. (ELIZA: people do not talk to the computer, but to the programmer who created the program.)

Explanation 3: Although humans do not **believe** that technology can be an autonomous source (and that social rules can be applied), they still **behave** like it is one (they still apply the social rules to technology)! Beliefs do not necessarily guide behaviors.

Addendum: Certain cues make humans apply social rules more easily, e.g., if a source provides a request and at the same time a justification (the cue). The justification does not have to be meaningful, can be even tautological!! ("May I please use the Xerox machine now? I have to make copies.")

Five cues to social responses

If technology provides certain social cues, humans will accept technology as an autonomous source ("self", "other"), even if it shows no motivations and attitudes. Such humans do not have to be abnormal, and they do not have to intend to interact only with the creator of the technological source. The cue is enough to make the human apply social rules to technology.

I. **Language**. Interaction based on language vs. based on numbers or images.
II. **Interactivity**. Reaction of technology is based on multiple prior inputs (context sensitivity)
III. **Filling social role** (teacher, doctor, tutor). Role is actor with certain behaviors.
IV. **Human sounding speech**. Speech processing is different from other acoustical processing.
V. **Human-like face**. Same as with speech: we perceive faces very specifically.

Modern computers have all that: word processing, word queries, context sensitive, fill social roles (telephone operators, bank teller, teacher, therapist, secretary, co-worker (robotics), game player, friend), animation and speech output. Therefore, such technology can elicit social reponses from humans. ((I am not sure if digitized sound and animation is already enough. And how good really can computers fill in social roles? And I still can't ask my computer a complex question, unless I use the computer specific syntax. These social cues seem to be a bit vague!!))

Characterizing sources as "self" and "other" ((or rather "same" and "other"?))

Different voices on computers seem to make human think of different sources ((entities)).

2. How does A.I. look at concept of agent and agency?

Agents can mean a lot: from fundamental units of actions of any kind (performing single simple action, Minsky), to behaviorally complex animated entities [Maes].

Autonomous agents: act without explicit external guidance
Intelligent agents: perform automatic processes based on particular constraints
Character: in films and novels, *can't affect humans in the real world*
Agent: can interact with humans in the real world; they are *useful*

Believable character:
(a) Displays personality
(b) Interacts naturally with other characters
(c) Exhibits behavior that is consistent with explicable internal states

Believable agent:
(a) Its behavior must be based on continually updated representations of the user's beliefs and desires
(b) User must trust them that they can do their appointed task

In general, AI people focus on the **construction** of systems, not on the observation and analysis of the resulting behaviors (usually no generalizations). Also, they often focus on **low-level details of human behavior**, not developing well-defined criteria for believability ((Is this still true? Watch the Final Fantasy movie!)) He suggests that AI people pay more attention to social criteria.

3. Anthropomorphic responses to technology. From HCI

Agents and agency in HCI: two approaches:

(I) **Anthropomorphism**: Attribution that computer IS a human.
Currently dominant, from A.I., ontological.
Example: Turing machine. Opponents: Shneiderman, Turkle, Weizenbaum.

(II) **Ethopoeia**: Computer only exhibits human characteristics (which includes human-like responses)
Less well-known, user-centered view.
By Nass, Steuer, Tauber. Leads to computer as social actor (described in next chapter).

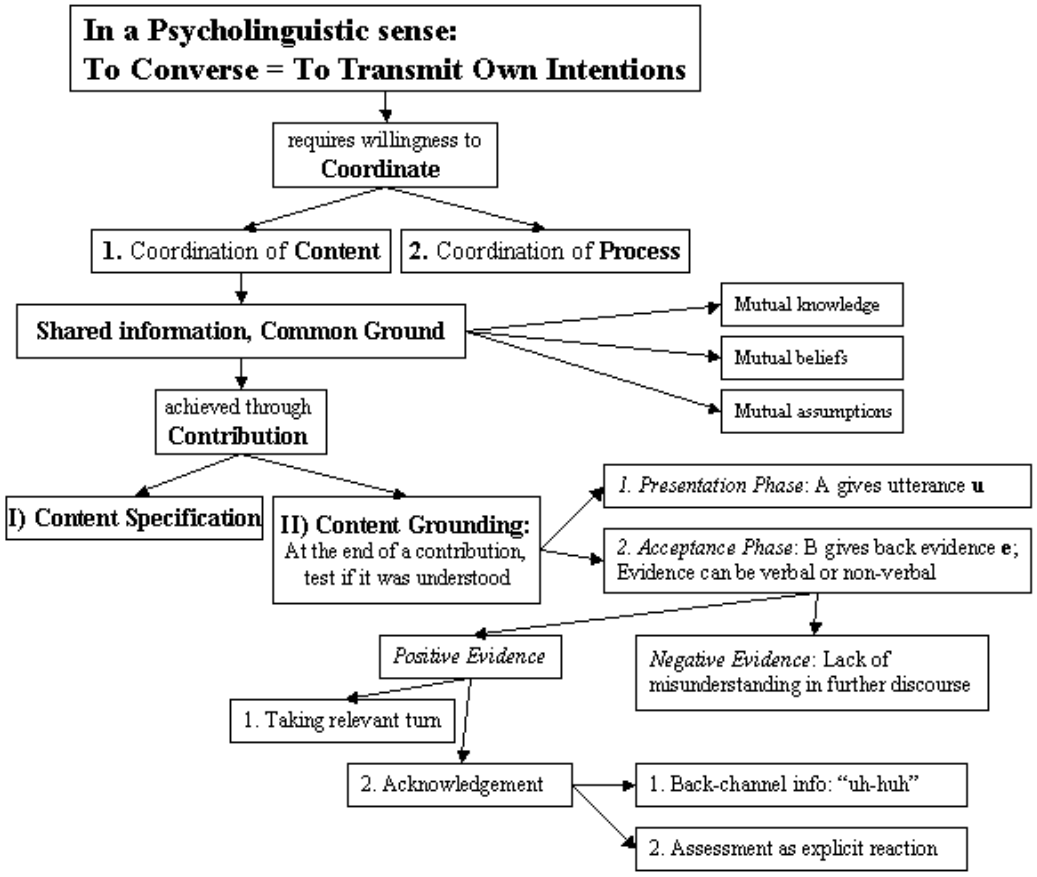
4. Can we use theories of conversation to design agents? From Psycholinguistics

Simulate human-human communications patters with computer agents. Trying to understand agency by analyzing conversations.

Example: ELIZA. Is very simple, but effective. Uses 3 of 5 social cues (language, interactivity, filling social role). That seems to be enough for humans to accept it as an autonomous source. ((Foner doesn't think Eliza is an agent, though!))

- In general, when people converse, they want to **transmit their intentions** to the others.
- This assumes the willingness to cooperate (Cooperative Principle, Grice), which requires **coordination**, a continuous adjustment to the situation and to the actions of the other party.
- Coordination: (1) of content (2) of process
- Coordination of content requires *shared information* or *common ground*, consisting of mutual knowledge, mutual beliefs, and mutual assumptions.
- Common ground is achieved by *contribution*: two-stage process: (1) *content specification* (2) *content grounding*
- By the start of a new contribution, the participants try to find out if both have understood that last one. This is also a two-phase process:
  (1) *Presentation*: A presents utterance "u" to B. If B gives evidence "e" or stronger, A thinks that B has understood what she meant by "u"
  (2) *Acceptance*: B accepts "u" by giving evidence "e".
- Positive evidence:
  (1) acknowledgements:
    (a) back-channel responses like 'uh-huh'
    (b) assessments that indicate the reaction
  (2) taking relevant next turn
- Negative evidence: lack of misunderstanding in further discourse
- Nonverbal cues (head nods, puzzled looks) can be both positive and negative evidence

Psycholinguistic background



Psycholinguistic background according to Steuer (my graphical representation of it)

Mediated interactions limit the modes of presentations: for presenting the common ground, and presenting evidence. Face-to-face provides the most possible means to present common ground; mediated has usually much less means.

Conversational strategies in the interface

When designing interfaces, one has to keep in mind common ground and grounding.

But in mediated interaction, knowing that the other party is a machine rather than a human effects the communication. If talking to computer, humans are more verbally concise and more focused (but it decreases over time).
Brennan and Ohaeri did a study with three message styles

(1) Telegraphic: sentence fragments, no self-reference in first person
(2) Fluent: complete grammatical sentences, no self-reference in first person
(3) Anthropomorphic: using "I"

Result: in the anthropomorphic condition, subjects used more words: "users' language use is shaped by the systems' language."

Adding human faces to computer interfaces can, but does not necessarily enhance the experience for users.
Thorisson used facial expressions and a conversational metaphor for designing computer-based agents ("J Jr.")

Conclusion

Four literatures:

- Relevance of source of message in general, and self-vs-other evaluation in particular (Social Psych, Sociology)
- Perception of technology as autonomous sources in HCI and AI, including "believable agents"
- Two schemes of believability: Anthropomorphism and Ethopoeia.
- Conversational situations, in Psycholinguistics.

Comments

Very useful and deep. Very nicely written. Nice approach comparing comments from **AI, HCI, Social Psychology, and Psycholinguistics**. The last one not very relevant for me though, since this knowledge would just help me if I want to re-create a human-agent conversational interface after human-human conversation, which I don't.

Authors say that computers have all the cues available to elicit social behavior from humans. I think that this is not true (yet).

Lars Oestreicher, Helge Huettenrauch, and Kerstin Severinsson-Eklund (1999). Where are you going little robot? - Prospects of Human-Robot-Interaction.

Summary

Introduction

Futuristic robots seem to interact naturally with humans. But we are not there yet. Mobile robots are a now frontier for interactive systems, because they differ substantially from other interactive computational systems.

The human-robot interaction group (NADA, Numerical Analysis and Computer Science at the Royal Institute of Technology in Stockholm, Sweden, http://www.nada.kth.se/index-en.html) studies:

(1) Task analysis for human-robot interaction ((moderately relevant))
(2) Multi-modal human-robot communication ((sounds interesting to me, but not relevant for Autonomy vs. Control))

Study 1: Questionnaire, assessing people's attitudes towards obtaining help from a domestic robot; 134 subjects.
Most of them were positive towards getting help from robots; females a bit less.
Important side result: 71% thought that robots would not invade their privacy.
How comfortable are they with robot's independence?

(a) Programmed robot: 78% positive
(b) Autonomous, and learning by doing: 48% positive
(c) Smart, take initiative and suggest actions: 56% positive

((Why are people more OK with "smart" than with "autonomous"? Is it an artifact of the questionnaire? Is smart more acceptable than autonomous?))

Study 2 (Khan, http://www.nada.kth.se/~zayera/report.pdf): Preferred way of communicating with a service robot.

Choice of modality:
(a) Speech 82%
(b) Touch screen 63%
(c) Gesture 51%
(d) Command language 45%
Combinations are also acceptable.

This study is also helpful in finding out
(a) Task analysis: the proper task that a robot should perform
(b) Communication and Modality: How to communicate with the robot in the best way

(a) Task analysis

TA for mobile robots is difficult, since tasks in household environment can be complex. Examples:
(a) If robot asked to pick up a bowl on the living room table, does it have to bring the popcorn in it too or not?
(b) If asked to bring a book, does it have to bring the other books that lie on top of it too?
(c) What should fetch-and-carry robot do if path is blocked? ((rather trivial))
(d) What should follow robot do if human stops? ((rather trivial))
Solution: Look at environment, be context sensitive.

A robot should be regarded as an end-user oriented computational mobile device and/or agent. The HCI community is asked to answer the following questions:
- What should a task analysis for intelligent service robots capture?
- Can the standard methods be applied?
- How to describe the context of a robot?
- What is specific to human-robot interaction that will complicate task analysis?

Problem with classic TA methods: They often focus on highly interactive tasks with close interaction, where the tasks are internal in a computer. Not so with service robots: they have to be environmentally aware.

Usage: How interactive can a robot be, how much can a robot ask the user for feedback before she gets pissed?
Trust: User trusts robot with autonomous tasks if it seems to be acceptable (probably: up to the challenge).

(b) Communication and Modality

The above-mentioned study suggests users prefer speech in combination with other interaction modalities.
Questions:
- Can user expectations be met with today's technologies? ((bad speech reco, etc.))
- Are methods of Natural Language Interfaces (NLI), multi-modality, and agents sufficient for interactions with robots?
- Do we need new heuristics?

Can we use existing human skills like face-to-face communication, mediated human-human communication?
Physical embodiment of robot requires might require new dialog strategies. Example: "Go left". Does it mean from user's perspective, or from robot's perspective? Robot has to detect the ambiguity, and/or initiate appropriate dialogue.

One solution would be to use multi modality: highly redundant, increased accuracy, possible synergy effects. Authors refer to Put-That-There (Bolt).

Questions:

(1) What should be the **principles guiding the design of interactive systems like mobots**, and how can we minimize intrusive systems like data-glove and eye-tracking that are usually used for multi-modal interaction? The authors think that *Asimov's robot rules* might be a good starting point.

(2) Is it becoming more desirable to **communicate with machines** than **operating them**? (Koons) What metaphors are appropriate for what kind of robots?

(3) *Concrete*: A fully trusted and autonomous robot will have **what heuristics**, **what range**, and **what communication and interface design**? If it acts as a social agent, would it have an **illustrative interface** or should it **interact only with the primary user**? E.g., can it leave voice messages and send email to a third person in certain cases? ((Good! Now somebody has to answer all these questions!))

Conclusions

HCI gets a new challenge with HRI (human-robot interaction) in domestic mobots, because of additional factors like
- Dynamic environment
- Object and context recognition
- Physical environment.

HCI might have to develop new methodology.

Comments

Moderately relevant, and not very well structured. Asking a lot of (good!) questions, but not answering a lot. But that's probably the idea behind a position paper.

Nevertheless, it seems like a collection of somehow related research themes, but I don't really see how Task Analysis is so relevant in this paper--perhaps because Lars Oestreicher is familiar with it and writing his PhD about it?

Valentino Braitenberg (1984). Vehicles: Experiments in Synthetic Psychology (book, 155 pages, get overview)

August xx, 2001.

Summary

Comments

I have read the first few chapters two years ago, and I sort of know the basic idea behind the vehicles.

Katherine Bumby and Kerstin Dautenhahn (1999). Investigating Children's Attitudes Towards Robots: A Case Study.

Summary

Short: HCI studies human attitudes towards software and UI. CT (Cognitive Technologies) extends this to robots, to HRI (Human Robot Interaction). **Autonomous agents research findings--which themselves come from human-human interactions--can be applied to robots. Currently, robots are built mainly for technical efficiency.** But

they should be built to **optimize human-machine interaction**, so they have to **have social skills (human side of robotics).** Study: How do kids (a) perceive and (b) interact with robots.

Important conclusion: Currently, **robots are mainly associated with either machines** (factory environment) or **fictional characters** (movies). Both are focusing rather on confrontation than on integration. However, robots should be built with social skills and adapt to cognitive and social needs of humans, allowing more human type of interaction between humans and robots. IMPORTANT: Social control should replace conventional, instructional means of controlling machines. ((In other words: Kids are OK with autonomous robots, as long as they are nice to us and we can talk to them in a human way.))

Cognitive Technology (CT): bridging the gap between the technological and human. Attitudes are important indications about human cognition. Human attitudes towards computers have been studied extensively in HCI.
Results:
(1) General tendency to anthropomorphize computers
(2) Significant influence on social and cultural factors
(3) How to use these factors in interface design (for software and UI)
But: Studies of human attitude towards technology should not only include software and UI, but also robots, as tools (rehab robots) and as toys (furby).

Currently, most research about robots is done on the technical aspects, but more should be done on the following aspects of robots:

- Appearance
- Believability
- Social skills required

Goal: Robots with human side.

Structure of paper:
(I) **Discussion of human machine relationship in HCI**
- User cognition
- Personality

(II) **How to build believable agents (biologic, software, robotics)**
- Attribution of social characteristics to machines
- Agents and personality
- Agent believability

(III) **Study**
- Aims and hypotheses
- Design
- Participants, Procedures, Results, Methodology, Discussion

User Cognition

Important for humans is to understand the behavior of others, and to infer the internal states ((like conversations are mainly for transmitting intentions?))
Process (Baron and Byrne):
(1) Categorize behavior, either
- for new events: is it internally or externally caused?
  Much cognitive processing required! Or assimilate schemas.
- faster: universal judgments: attitudes and stereotypes,
  both are schemas; can be on unconscious level! (Nass)

(2) Infer character traits (characterize behavior to infer traits)
(3) Correct inferences with situational information
What schemas to people have about robots? Probably not many. ((Disagree. Many schemas from Sci-Fi movies and toys))

Personality

Usually measured in 5 traits: extroversion, agreableness, conscientiousness, emotional stability, and openness to experience.

Interactions:
(1) Similarity of personality
(2) Complementarity of personality
(3) Oppositional personalities (if one has undesirable personality traits, also economical status and physical attractiveness.)

Attraction between personality types can come from both (1) and (2) (depending on whom you ask).
Notes:
- It is not only about assessing the *other* person's personality, but also the *own*.
- It is not only about *actual* personality, but also about the *perceived* one.

For building agents:
- Which roles/tasks is it assumed to perform?
- What relationship model does the user prefer, (1) or (2) or even (3)?
- In general: give the robot positive personalities ((duh))

Attribution of social characteristics to machines

Nass says: If machines (computers?) have personality like characteristics, people will respond to them as if they have personality.
People are sensitive to dominance level of computer, and affiliation and competence is linked to that (like with humans).
People also prefer interactions with computers with similar personality to themselves.
People can treat computers as social actors (sources of messages), even if they know that it is inappropriate: contrast between belief and behavior (same as with humans).

Agents and personality

How can personality traits be expressed in robots? ((Ken Perlin: mobot movements map to gestures??))
- Nass: Certain isolated behaviors (of computers) can cue the attribution of certain character traits.
- Dryer: People prefer same characteristics as in humans (cooperative, outgoing, calm, organized, curious), same characteristics as they have, and strongly expressed characteristics.
- Dautenhahn: Dangerous to transfer personality preferences for other humans to agents. Some character traits that are positive in humans might not be positive in robots, e.g., "socially proactive": persistent/pushy (human, pos/net) vs. intrusive/scary (robot).

Gender: Nass: Even minimal cues prompt male/female and masculine/feminine stereotypes for agents (same as for humans)

Appearance: Humans respond more positively to attractive people. Humans prefer those who have same level of attractiveness. Goodies are attractive, baddies are ugly. All applies to agents. Therefore, appearance of robots should match their function.

Important Note: all personality perception is confounded with cultural factors. Therefore, agents have to adapt to the culture they are in.

Agent believability

Believable agent = provides the illusion of life (Bates).
Important: Believability is in the eye of the observer, not a property of the agent itself.

How to make them believable?

Nass: copy human interaction behaviors. Sophistication of copy can be minimal, as cartoonists know.
Believability can take the form or anthropomorphization: if agent is anthropomorphized, it is believable. Important: still many problems with creating anthropomorphic characters.
Note 1: Cues to anthropomorphize can be minimal.
Note 2: Tendency to anthropomorphize could be innate.

Study

How do kids perceive robots, and what behavior do they show when interacting with them?

38 subjects. Drawing a robot and writing a story about it (as a single), and the interacting with two robots (as a group).
Robots: Braitenberg vehicles with "fear" and "liking" behavior, and bumping sensors.
Interaction: See how the program gets downloaded, touch the robots (bumping sensors), move the light source.

Hypotheses and questions:
H1: Kids will tend to show robots in social settings
YES. School or family setting

H2: Kids will tend to treat robots socially
YES. Treat them like pets. But also non-social explanations, due to fluctuating levels of believability.

H3: Kids will tend to place the robots in familiar settings (performing similar tasks, with familiar people/tasks)
YES. Robot that plays football.

H4: Kids will tend to anthropomorphize /animate the robots in their pictures and writing
YES. Drawings with humanoid faces and feet (!), writings with robots with emotions and preferences, etc.

H5: Kids will tend to anthropomorphize /animate the robots even if there are no intentional cues (study c)
YES. Talked to robots, etc.

H6: The personality of the kids will affect the level and nature of their interaction with robots (study c)
Looks like. Age and personality had influence on interaction. ((But no personality tests done)) Kids of age 3 and 6 were not interested: the former because too young, the latter because technology too trivial.

Q1: Will the robots be perceived as believable?
The robots are very low-level sophistication (Braitenberg vehicles), but it should still work) As said above, yes for age 4 and 5, no for 3 and 6. ((Interesting. Kids would think these robots are believable, but only for two years in their lives?? That can't be right.))

Q2: Will the robots be portrayed as violent? (Terminator, etc)
No. Some did, but they were all male, so it must be more a trait of subject gender than a trait or robots in general. Also the language was violent, but not the robots themselves, but what happened to them. The authors say that kids naturally respond like that. ((This needs further examination. If an entity attracts violent actions, isn't partially because of this entity?))

Q3: Will the kids tend to attribute gender to the robots?
With drawings no, with stories and in interaction yes.

Q4: Will the kids tend to attribute specific gender to the robots?
Yes, male. It could be that technology in general is male, not robot specifically. We also have to find out if robots in fiction are mainly male and if kids are influenced by that. ((This is true for almost all questions asked here!!!))

Q5: Will the kids possess a stereotype of a robot?
Yes. Geometric shapes, humanoid face and feet. Social, non-violent, male, some evidence of free will.

Kids tend to place robots in positive schemas, not related to robotics. Perhaps in future create robot specific schema? ((Huh? Kids have seen robots in sci-fi for a long time. Of course they have a robot schema.))

More findings:

- In drawings, robots have **geometric shapes**
- In stories, robots have **free will**.
- ***Very Important***: Also in direct interaction, kids attributed free will, *even when they saw the program being downloaded*. ((DO they understand the concept of downloading??)) This means that the kids do not think that they can control them anymore. This is very relevant because today, the feeling of control over agent/robot is considered the most elementary requirement. **This research shows that this is perhaps not justified: control might not be so important, if the robot supports high-level interaction**. ((I like that conclusion. It means, if I can talk to a robot in a human way, high-level conversation, then I can accept that I am not in control anymore. Would this mean that if robots get to a level of sophistication that we can talk to them in a natural way, we accept them being completely autonomous?))

Comments

This is the description of a survey and how they came to it. The methodology of the study is fuzzy, at best. The more interesting part though is the first, where they describe Personality, Attribution of social attributes to machines, agent believability and personality, etc, and the conclusions.

The most interesting finding is that kids attribute free will even when they see how the robot is programmed. The authors go a bit further and say, that they are willing to attribute free will *if it is possible to interact with the robot on a high level* (where do they get that from??). Anyways, I like that conclusion. It means, if I can talk to a robot in a human way, high-level conversation, then I can accept that I am not in control anymore. Would this mean that if robots get to a level of sophistication that we can talk to them in a natural way, we accept them being completely autonomous? Is sort of implied by Nass's finding of treating machines as social beings when they have language, human voice and face, assuming that being a social being includes autonomy.

Nice conclusion paragraph: Robots have to be built to be social, and that means we have to be able to interact them in a human way, and social control should replace conventional, instructional means of controlling machines.

Kerstin Dautenhahn (1998). The Art of Designing Socially Intelligent Agents - Science, Fiction, and the Human in the Loop.

July 22, 2001

Summary

Abstract

In this paper, **SIA** should not only **behave socially** (from observer point of view), but also **recognize and identify other agents**. So the question is: how to design SIA with a human-centered stance. This research is part of the EAL (embodied artificial life) research.

Cognitive Technology (CT) approach towards SIA:
(1) New forms of interaction at the human-tool interface.
(2) How social agents can constrain their cognitive and social potential.

(3) How SIA technology and human social cognition can co-evolve and co-adapt: this requires a cognitive fit.

SIA might be most successful if they are a bit like us: Human social psychology can help to create a believable agent:
(1) Storytelling
(2) Empathy
(3) Embodiment
(4) Historical and ecological grounding (former: autobiography)

These concepts should avoid the shallowness of approaches that only take advantage of the anthropomorphizing tendency in humans.

Introduction

Definitions of agents: common sense, formal, computational.

1. Common sense (from dictionaries)

- Person who acts
- Person used to achieve something
- Agents are persons

Therefore:
A. Agents have purpose/goals
B. There are several agents
C. They interact with each other

2. Formal [Luck/d'Inverno]

- Objects = having set of actions and attributes
- Agents = objects with goals
- Autonomous agents = agents with motivations that generate goals

Therefore:
(1) Agency works with single agents
(2) Agency is transient, not internal, *attributed by observer*
- Object can't "be" an agent, but is "interpreted" as one
- Environment "affords" agents
(3) However, "autonomous agency" is encapsulated, does not need to be attributed, since goals get generated internally!

3. Computational [Franklin/Graesser]:

"Autonomous agent = system situated in and part of an environment that senses the environment and acts on it, over time, following its own agenda, and so to affect what it sense in the future."

This definition includes robotic, computational, and biological agents. ((AM would be such an agent.))

AI, AL, EAL

SIA is part of
- AI: systems that ARE intelligent (strong AI) or BEHAVE intelligent (weak AI)
- AL: systems that behave life-like.

BUT: Dautenhahn argues that "intelligence" and "life" can not be defined, but are both constructed and attributed by humans, and therefore subjective! It depends on (a) the *context* they are located in, (b) on the how humans *agree on the interpretation*, (c) and *social conventions* in the environment of the observers.

In general, there are two AL approaches:

(1) Old fashioned AL: quest for the logic of life
(2) Embodied AL (EAL): "studying the natural form of complexity in artificial media by constructing systems" ((what is THAT??)) "Constructing systems based upon the physical properties of the matter, not abstract formalisms."

EAL is basically AL with CT approach.
CT = study of integrative processes that shape human-object interactions. How technological tools:

1. Influence human perception
2. Affect natural human communication
3. "Act to control human cognitive adaptation" ((?))

Therefore, to reach cognitive fit between humans and the technological tools, one has to understand human perception, communication, and social and affective constraints. But, the human-tool interaction does not have to mimic nature and copy "natural" forms of interaction, it can be different. Some kind of "Interactive Intelligence" could emerge that is more than the sum of its parts (human plus tool): it would be a "dynamic spatio-temporal coupling between systems, embedded in a concrete social and cultural context." This kind of intelligence is different from the classical intelligence that is usually solely a property of the system itself.

History of autonomous agents

Dautenhahn uses an extension of the autonomous agent definition by **Franklin/Graesser**:

> "Autonomous agent = system situated in and part of an environment that senses the environment *and other agents of the same and different kind* and acts on it, over time, following its own agenda, and so to affect what it sense in the future."

- Agents are located in a habitat.
- They use the habitat's resources, and therefore change their environment.
- They have full or partial autonomy over energy and space.
- The first autonomous agents were biological; they exist for 3.5 billion years now (single cells, multi-cellular organisms, animals and plants).
- Homo (2.5 millions years) and Homo Sapiens (100,000 years) are special because of increased interaction with other agents (as well as manipulation, representation, and control). They are mainly social animals, and had to cope with complex social relationships.
- The "**Social Intelligence Hypothesis**" claims that human intelligence originally evolved to solve social problems, and only later was extended to problems outside the social domain (math, abstract thinking, logic)

- Individualized vs. anonymous societies:

a) Ant/bee/termite colonies are anonymous societies, where the individual plays no crucial role. "Ants don't have friends" and "Ants don't tell stories": they interact through the environment: stigmergy.
b) Humans developed individualized societies. This means, they bond socially, have attachment, alliances, dynamic hierarchies, etc. This leads to **primary groups** which consist of family members and close friends (not necessarily genetically related); they can be test beds for social behavior. Dautenhahn wants to generalize the concepts of primary groups and individualized societies from biological to artificial agents.

The self of a member of an individualized society can be defined by **autobiographic memory**: it remembers what happened to it and re-interprets that history continuously.

Human societies were growing to about **150 members** (max number of people that one person can maintain a stable relationship with only face-to-face relationships). To get that far, they developed *social grooming* techniques: language that allows communicating on higher levels of abstraction than body language and facial expression. Therefore, even today, humans use language to communicate stories about other (gossip, 60%). ((It glues a society together.))

Furthermore, **humans seem to be "naturally biased to perceive self-propelled movement and intentional behavior":** we naturally animate and anthropomorphize. This is true not only for biological agents, but our self-made computational and robotic agents. Humans naturally want to interact them in a natural, human way. We want them to be believable. Such agents can be viewed as a new species that should be integrated in our society.

This leads to the question if it is desirable to construct *social interfaces* that bridge agents of different species. Within the community of computational and robotic agents, the communication does not have to be human like, though. But with humans, they have to interact naturally, in a way that is acceptable and comfortable to humans, so that they get accepted as 'interaction partners'. Research is done in this area, called: 'human-robot symbiosis', 'mixed-initiative problem solving', and 'co-habited mixed realities'.

Human social intelligence

Studied in many fields like psychology, sociology, ethology, economic. Here only focus on four aspects:

(1) Embodiment
(2) Empathy
(3) Autobiography
(4) Narrative agents

Embodiment

All biological agents are. Question is not IF embodiment matters, but HOW.
- Robotics: New AI is Embodied AI (EAI) (Brooks), which leads to cognitive robotics.
- Computational: What does embodiment mean here? Examples:
  - Tropical rain forest vs. memory space in computer?
  - Sensory-motor systems of animals vs. input-output functions of a program?
  - Flock of migrating birds vs. mobile software agents on the net?

Scientific discussion on finding the right level is still open; danger of ending up with purely metaphorical comparisons!!

Empathy, or rather memory and stories

- Paradigm shift in AI: Psychology says that human understanding and interpretation is based on stories.
- [Schank, Abelson] Scripts as representations of generic event memories; scripts as suitable computational approach towards story-building systems. New experiences are interpreted in terms of old stories. "Remembering static facts (phone numbers) are the result, but not the basic units of remembering processes." Remembering = creating and inter-relating stories. Stories enable us to deal with the complexity of social interaction.
- Problem with scripts: capture only generic, canonical behavior in a culturally defined situation. They are abstract data structures, abstract away from the individual. What is missing is one's personal life history, the autobiography.

Autobiography

- Autobiographic agent = embodied agent that dynamically reconstructs its individual history during its lifetime.
- In AI, consistency is important. Humans don't care: "the subjective impression of being a static personality is an illusion." Believability is more important than consistency.
- Biological agents can only be understood with reference to its history, both evolutionary (phylogeny) and individual development (ontogeny).
- Humans are constantly telling and re-telling their stories, they are autobiographic agents.

Stories about oneself and others

- Telling autobiographic stories about oneself and biographic re-constructions about other persons is the central set of mechanisms that contribute to "social intelligence."
- Humans are addicted to stories.

Agent technology from the observer point of view

Believability

- Believability is in the eye of the observer, influenced by the observer's personality, naive psychology, and empathy mechanisms. Therefore, there are on 'objective performance parameters.'
- Examples: Luxo Jr. and Toy Story. The former does not mimic the morphology (structure, from) of any specific species, and still it can show life-like properties. (Cheap just to add a tail and big eyes...)
- "Believable agents are said to be scientifically 'cheating', because they put all the intelligence in the human-agent interface, and rely on the intelligence of the human using and interpreting this interface." True, but it is only bad if one assumes traditional AI that tries to put intelligence into the agents themselves. New AI has not problem with that.
- It is more useful to look at the individually constructed reality than an objective reality. The meaning, and not the technological basis, is central.

Virtual pets as believable agents

- Are life-like: living in an environment, can express emotions, can die, can interact with humans. They don't LOOK life-like.
- Critics say that they are cheating, pretending to be more intelligent that they actually are, or they 'exploit' natural human instincts of nurturing and caring.
- Autobiographic feature can increase the believability: some show 'ontogeny' by developing into different characters: simulated growing up, showing biographic history.
- Cyber pets are not complex or intelligent. They just exhibit interesting behavior during interaction with user.
- Interactivity makes cyber pets believable and popular.

Believable fictional biology: Rhinogradentia

What does this story teach us?

1. Life-like artifacts need not necessarily mimic nature, but it helps if their appearance and behavior are biologically plausible. There can be alternative life forms.
2. 'Form fits function', and 'function fits environment.' Social agents become more plausible if embedded in a rich and plausible story. The more we know about life and autobiography, the more believable.

Social agents within and Embodied Artificial Life framework

Sciences of the artificial

- AI: human intelligence is/can be modeled as computation. Focused on human problem solving, neglected development (ontogeny and phylogeny), embodiment, personality, creativity, social skills, emotional skills. Had influence on what human intelligence was supposed to be: logic, mathematical skills, spatial thinking, planning, etc.
- AL:
    - Bottom-up-approach: systems are understood by synthesis (not analysis)
    - Based on local non-linear interactions between components, the emergence of complexity on the next higher level is studied
    - Emergence cannot be predicted: either we let it run, or we apply evolutionary techniques

Two AL research directions (not the same as strong/weak AI):
(1) Finding the 'Logic Of Life' (similar to classical AI): what are the objective criteria for life?
(2) Creative story-telling approach: "Find forms of complexity in artificial media which appear to be natural, and which give us plausible explanations about life." ((??)) No objective criteria, definition of life is in the eye of the beholder, every individual seems to have his/her own conception of life. This is usually EAL.

About (Not) Modeling the Social World

It is not necessary to model the social world. Social behavior will emerge from situated activity.

- "How to make agents social" seems to be parallel to "How to make robots intelligent": classical vs. behavior-based robotics; building up a model of the external world vs. interacting with the real world through online real-world sensor data.
- ==>> **"The world is its own best model"**
- Behavior-oriented robotics seems to be related to behaviorism: not true, because it does not doubt the existence of internal psychological processes. It only rejects the idea that these are based on symbolic, internal world models (as classical AI suggests)
- Many behavior-oriented approaches toward robot sociality focus on stigmergic communication (through the environment), like insect societies.
- Examples for emerging behaviors: turn-taking behavior in mother/kid relationship.

SIA and Human Nature

Not all social behavior is good, e.g., violence. We don't want agents to have this one.

- SIA should learn from Human Factors (ergonomics) about how to design technology that works well with humans.
- BUT 'humanizing' the interface depends on what one includes in human!
- Although violence is important to life and survival, but we don't want agents to be violent.
- SIA agent models can get inspired by human social behavior, but we have to make explicit which model of sociality we refer to; and there are also cultural differences.

Summary

- **Embodiment**: structural and dynamic coupling of agent with environment; both physical and internal
- **Autonomous agents**: agents that inhabit our world, react and interact with their environment and with other agents (same or different kind) (extended Franklin/Graesser). Can be biological, computational or physical agent. Biological agents can only be understood with reference to the historical context, their phylogeny and ontogeny; aliveness, autonomy, and embodiment are interconnected.
- **Believable agents**: agents that are presented to humans as 'characters', vs. 'intelligent agents' that act in the background. If they appear life-like, humans find them interesting and can develop personal relationships. (Biological agents are genuinely believable, since they ARE alive.)
- **Autobiographic agents**: embodied agents that dynamically reconstruct their individual 'story', with dynamical memory. Reflect stories about themselves and relationships with other agents.
- **Human social intelligence**: live in individualized societies; human style social intelligence means to develop and manage relationships between individual agents (more)
- **Socially intelligent agents** (SIA): any kind of agent that shows human style social intelligence; individual of non-individualized society (e.g., insects) is not an SIA.

Conclusion

Design guidelines for SIA, according to EAL approach (CT perspective on AL):

(1) **Humans are embodied agents**. Therefore, SIAs should be able to handle both objective and subjective time in human dialogues and in the way humans remember events and personal experiences.
(2) Humans are **active agents**, want to use their body and explore the environment. Therefore, the more degrees of freedom an interface has, the better for the human.
(3) Humans are **individuals**, and they want to be treated as such, even if they have the same genotype. Therefore, developing individuality is important for SIAs: Imitation and social learning make agents more like us.
(4) Humans are **social beings, not problem solvers**. Human intelligence can develop and be understood only in social context. Since we are not mainly problem solvers, tools can assist us in these areas, reducing cognitive load on daily routines (e.g., mail filtering)
(5) Humans are **storytellers**. Creating and reconstructing stories is crucial for human understanding. SIAs have to be good at telling and listening to stories; that can be text, but also pictures, or non-verbal communication.
(6) Humans are **animators**. They tend to animate and anthropomorphize the world, add intentionality. To exploit that tendency, one could build agents that copy nature (humanoids), with the disadvantage that the expectations might be too high. The other possibility is to design believable systems without mimicking nature.

(7) Humans are **autobiographic agents** and **life-long learners**. They constantly learn and re-learn, re-write their autobiography. Systems that help humans to re-construct autobiographical memories strengthen social skills and the self [Rosebud, Jennifer Glos]. The 'social interaction hypothesis' says that autobiographical memory develops gradually: by talking about memories, kids learn to structure and retrieve them.
(8) Humans are **observers**. Human perception and cognition is subjective. Human behavior and motivations can only be understood in historical and cultural context (e.g., Aztec cannibalism). Agents have to adapt to cross-cultural differences as well.
(9) A matter of balance.
(a) Cognitive fit between agent and human
(b) agent has to be historically grounded (autobiographic)
(c) believability needs complete design: social, ecological, and cultural situatedness of the agent and its body.

Applications for SIA:
- Fun context: games, virtual pets. User probably chooses the more believable product that allows the most natural social interaction. Or: personal assistant that cares for a single user and shows human-style form of social intelligence.
- Business context: functionality and efficiency are relevant. Expressiveness and believability have to be relative to the functionality: a SIA is not appropriate if it requires too many resources (hardware, user's attention and cognitive load) to complete the task!

Therefore, there is a **tradeoff between 'efficiency' and 'sociality'**. SIAs are appropriate if their function is primary social. They should be used when personality, character and individual relationships are desirable.

Our societies are changing, and also our conceptions of 'being social'. SIAs might be a mirror of our own sociality. ((Are we really changing? I don't think so.))

Comments

In general, the paper is very badly organized: the titles and sub-titles are sometimes not accurately describing the content of the chapters. She repeats herself a lot. There is no clear line of thought, but a collection of loosely related concepts and ideas.

Her alternative approach to AL is strange: What are "artificial media"??

Look up HUDL:
- http://www.google.com/search?q=cache:J9g40xuWZ1A:shogun.vuse.vanderbilt.edu/CIS/IRL/html/papers__1997.html+r.+t.+pack+hudl&hl=en
- http://frontweb.vuse.vanderbilt.edu/vuse_web/eng2/capbriefs.asp?BriefID=36
- http://www.engnetbase.com/pdf/0339/Chapter11.pdf

Remarks on Asimov's Three Laws of Robotics. http://www.androidworld.com/prod22.htm

July 22, 2001

Summary

1. A robot may **not injure** a human being, or, through inaction, allow a human being to come to harm.

2. A robot must **obey** the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must **protect its own existence** as long as such protection does not conflict with the First or Second Law.

Problems with *first law*:

- It **excludes robots from functioning as a soldier** (good), **policeman** (dunno), **security guard** (not good). Policemen are normal human, so they have to protect themselves BEFORE anything else, so they more likely shoot before they should. Robots don't have to do that, that's why they would be better cops ((hmm)). Robots are more suited as security guards because they do not get bored, tired, or distracted. ((Yes, if they are as good and diverse in understanding a situation as humans otherwise.))
- It also doesn't allow a robot to **defend itself**, especially against android hating people. But the android is my property, was expensive, and I might have an emotional connection to it. Therefore, it should be allowed to defend itself.
- In some situations, **both action and inaction will result in harming people** (e.g., two people fighting, and androids trying to stop them)

Problems with *second law*:

- It doesn't work because it is **not specific enough**: other humans could hijack my android. I couldn't rely on my robot because other people could give it new orders. Furthermore, it could be given violent, illegal, or immoral orders. First is rather easy to avoid; second: perhaps, but not very likely soon; third: needs a lot of work before a robot is ethically aware!

Problems with *third law*:

- Is sort of **obsolete with a bit or common sense**. Usually humans do that, but only to avoid injuring themselves with stoopid errors. Robots would not do that in the first place. Furthermore, they sometimes have to protect their owners actively with their lives (owners defined in the startup procedure).

Comments

Is it possible that **these laws get obsolete with some commonsense knowledge**? If the robot has the correct "self image," e.g., as a servant to a limited amount of people, then it would behave as expected. However, the function of the robot in society is the crucial part. Can robots protect the owners against other people? Against other robots? What is the worst-case scenario in a mixed human-android society?

Cynthia Breazeal (1999). Robot in Society: Friend or Appliance?

July 23, 2001

Summary

Abstract:

- Influence of synthetic emotions on human-style human-agent interaction.
- Present control architecture that does that.
- Show how Kismet works.

Robot in Society

- There are more and more robots, getting more and more complex, interacting with lay people. Therefore, robots have to be developed that interface properly with untrained humans: intuitive, efficient, and enjoyable.
- **Socially Intelligent Autonomous Robots** (SIAR) do that: they are not only about transferring task-based information via intuitive communication channels, but address the emotional and inter-personal dimensions of social interaction.
- *Example*: Kismet.

Robot Appliances

- To design good HRI, human factors have to fit: **how are people naturally inclined to interact with and user these technologies, and how does the use impact the person?**
- Classically, use natural speech and gesture to train and supervise the robot.
- **Challenge: Give the robot enough autonomy to carry out its task and still respond to commands of humans that oversee its performance**.
- One solution: Make robot learn a lexicon (mapping from sensory perceptions to linguistic symbols). Gradually teach a robot a more sophisticated instruction set

Socially Intelligent Robots

But there are other kinds of robots that do not just perform tasks. Some applications require a more social form of HRI. E.g., interactive animated software agents.

The more interactive these agents are, the more believable people want them to be. Animation characters are one example, pets another for anthropomorphization.

A Question of Interface

Dautenhahn suggests design considerations for SIA. We concentrate on four:

- **Human perception of SIARs**: How do people perceive SIARs? How do these perceptions influence the way people interact with SIARs?
- **Natural communication**: What channels of interaction are the most natural?
- **Affective impact**: How to interactions with SIARs impact people emotionally?
- **Social constraints**: what are the constraints in human-style social interaction?

Human perception of SIARs

- [Dennett] Own and other's behaviors are interpreted in terms of intentions, beliefs, and desires.
- Therefore, SIARs should be able to convey intentionality: they don't have to have them, but the user should be able to predict and explain the robot's behavior.
- In general, classical animators are very good at that. However, Breazeal thinks that it is doubtful that superficial mechanisms of animation can be scaled to unconstrained social interactions between humans and SIARs. Expressive acts eventually have to be generated by some kind of synthetic emotions.

Natural communication

- For task based interaction: speech and gesture.
- For more social interaction: perceiving the other's motivational state (beliefs, intents, wishes) is important, with
  (a) Affective cues: facial expressions, prosody, body posture
  (b) Social cues: gaze direction, nods of head, raising eyebrows, etc
- Also important to regulate the rate and content of information transferred (slow down, repeat)
- SIARs must not only send these cues, but also perceive them.

Affective impact of SIARs

- People anthropomorphize pets, emphasize, bond, especially if they respond on an emotional level (wagging tail, etc). Similar with agents, if they would be believable.

- But we are not there yet! Today: how to design SIARs that are not annoying or frustrating to users. Wrong expectations are shaped by SIAR appearance. Therefore, good designs make the user interact with the robot at the right level of sophistication.

Social constraints

- "Humans expect to share control with those whom they socially interact with." (That's the difference between interacting in social vs. physical world): turn taking, shared attention.
- Social exchanges between people are mutually regulated. Each participant responds and adapts to the other.
- They predict and socially influence each other's behavior.
- Prediction works only because the participants are cooperative, open to communication, and respect social norms.
- SIARs might need mechanism for empathy. An empathetic mechanism might be necessary for teaching SIARs the *meaning/value of their actions to other and society*. They must be able to evaluate the "badness" or "goodness" of their actions! That would help them to discern socially acceptable from unacceptable actions. (Note that that might be culture specific!) Some of the common sense knowledge ((??)) can be pre-programmed; some of it has to be learned by the agent.

Synthetic emotions

- Emotions are important for social interaction.
- In constrained scenarios, the designer could use just the anthropomorphizing tendency of the human.
- In unconstrained scenarios, SIARs must be able to express and perceive emotions: that needs
    - Synthetic emotions ((really??))
    - Empathetic learning mechanisms
    - Affective reasoning capabilities

((Does this all include common sense reasoning, or moral reasoning??))

Architecture of SIAR

Kismet is an example SIAR. Located in a benevolent environment.
- Task: to engage people in face-to-face social interaction, and to improve its social competence from these exchanges.
- Scenario: robot child playing with human caregiver.

Attention and Perception

- Like kids, Kismet has three basic low-level feature detectors:
  (1) Face finding
  (2) Motion detection
  (3) Color saliency analysis
- These three, combined with habituation, and motivations/drives/emotions influence the visual attention system.
- Perceptual stimuli selected by the attention process are social (people = move and have faces), or non-social (toys = move and are colorful)
- Percepts are combined into a set of *releasing mechanisms* ((??))
- Each releasing mechanism is tagged with arousal, valence, and stance values. Somatic marker ((??))

Motivational system

Consists of **drives** and **emotions**:

Three drives:

1. **Social drive**
2. **Stimulation drive**
3. **Fatigue drive**

- Each drive has a desired operation point, and acceptable bounds: *homeostatic regime*.
- Each drive contributes to valence and arousal measures.

Current emotional state:

- A point in a 3-dimensional space (affective space):
  (1) (1) Arousal (high/low): surprised--tired
  (2) Valence (positive/negative): happy--sad
  (3) (3) Stance (open/closed: accepting—defensive

- Current emotions are computed from current values of drives, releasing mechanisms, and behaviors. Kismet is always in a distinctive affective state at a given time.
- Affective space is segmented in *emotion regions*, e.g.

	Valence	Arousal	Stance
Happiness	+	neutral	neutral
Sadness	-	-	neutral

The emotion region that is closest to point in affection space is active. The closer to the region, the more intense the emotion (radial distance).

The motivational system influences (a) behavior selection and (b) attentional focus. The more intense the emotion, the more weighs a stimulus.

Behavior

- Behaviors are organized hierarchically. Three levels:
  - Level 0: task level; activated based on strength of associated drive.
  - Level 1: strategy decisions
  - Level 2: sub-task divisions
- On each level, there are **cross exclusion groups** (CEG) = competing strategies for satisfying the goal of the parent.
- Only ONE drive (the active one) can influence the robots face ((isn't this a bit restricting? I can be tired and wanna play at the same time!))

Expressive Motor Acts

- Level 2 behaviors evoke motor acts.
- Each emotion region (affective region) is mapped to an expression space: Kismet has six prototypical expressions, which span the whole space of all possible expressions.

Social learning

Ongoing development. ((Not clear to me what really works and what she plans to implement.)) "Kismet's affective states mirror those of the caregiver. These learning mechanisms bias the robot to learn and pursue behavior that please the caregiver (('goodness')) and to avoid those the displease her (('badness'))."

Social Interaction

Current architecture produces interaction dynamics similar to five phases of infant-caregiver social interactions:
  (1) Initiation
  (2) Mutual-orientation
  (3) Greeting
  (4) Play-dialog
  (5) Disengagement
These patterns are not programmed, but they emerge!

Her Summary

- When Kismet is bored, it successfully can make a human play with it with a toy.
- When Kismet is lonely, it successfully can make a human engage in a face-to-face exchange.
According to the four points mentioned above:
  (1) Its behavior conveys intentionality.

(2) Its facial movements and gaze are interpreted as social cues.
(3) Its affective displays in response to inter-personal exchanges has affective impact on user
(4) It can share the control with the caregiver to regulate the intensity of the interaction
- It is an emotion-inspired system that focuses on affective factors rather than perception and task-based behavior.
- It is a critical step towards socially intelligent artificial creatures that eventually might interact with us as friends instead of as appliances.

Comments

Goal is to construct a robot that is a friend instead of an appliance. She concentrates on four points and creates an architecture and a prototype that demonstrates the points.

Clifford Nass, Steuer, J., Tauber, E., and Reeder, H. (1993). Anthropomorphism, Agency, & Ethopoeia: Computers as Social Actors.

July 26, 2001

Summary

Abstract: Anthropomorphic responses to computers do not have to come from complex, agent-based interfaces, but from minimal social cues (such as human voice) that makes people apply social rules to evaluate computers. Different voices are perceived as different agents (but not different computers).

- Ethnographic and anecdotal research says that humans mimic human-human relationships in human-computer interaction. Since this is stoopid, these humans must be sick (ignorant, or psychological or social dysfunctional).
- To provoke such responses, it has been assumed that one needs complex agents with animated faces, use of language, use of first-person references, etc.
- The current research however suggests that we need only minimal social cues to use social rules towards computers. Users know and believe that computers do not have "selves", they still behave as if they would.
- Ethopoeia: assignment of human attitudes, intentions and motives to non-humans.
- Minimal cues they used are:
(1) Words for output ((language))
(2) Human-sounding voice
(3) Filling of social roles
(4) Responses based on multiple prior input ((context-sensitive))
- Social rules tested are:
(1) Praise of others is *more valid* than praise of self (accuracy testing)
(2) Praise of others is *friendlier* than praise of self (friendliness testing)
(3) Criticism of others is *less friendly* than criticism of self (friendliness testing)
- Tested were 88 CS students in 8 groups. It was a tutoring situation after which both the subject and the computer tutor were evaluated.
- Method: Subject was *tutored*, then *tested*. After that, the computer *evaluated* the tutoring session. Eventually, the subjects filled out a *questionnaire* about if the evaluator was friendly or not.
- Independent variables:
(a) tutor and evaluator: same voice vs. different voice
(b) tutor and evaluator: same computer vs. different computer
(c) tutoring session was mostly praised vs. mostly criticized by the evaluation computer
- Dependent variable: How friendly was the evaluator?
- Results:

- When the evaluating computer *praised* the tutoring, it was perceived as friendlier and more accurate when it had a *different* voice from the tutor.
- When the evaluating computer *criticized* the tutoring, it was perceived as friendlier when it had the *same* voice as the tutor.
- There was no difference for same vs. different computer for either accuracy or friendliness.

- Results short: Humans use some social rules in interaction with computers, especially when there seem to be two distinct agents based on distinct voices.
- Future research: WHICH characteristics of computers encourage WHICH users to use WHICH social rules when interacting with computers?

Comments

Why do computers appear super-friendly when they contain two agents and one praises the other?? Shouldn't two distinct computers be friendlier when one evaluates the other?

Kerstin Dautenhahn (2000). Socially Intelligent Agents and The Primate Social Brain - Towards a Science of Social Minds.

August 3, 2001

Summary

Abstract

- SIA in the context of how humans and primates interact with social world.
- Both phylogenetic and ontogentic issues are discussed.
- Implications for designing artifacts and evolvability of human societies.
- Theory of Empathy
- SIA research can help with science of social minds.

Introduction

- Reeves/Nass media equation applies to SIA.
- Design guidelines for SIA: Humans are embodied, active agents, individuals, social beings, storytellers, animators, autobiographic, observers.
- SIA research is more than just about software, it is rather part of autonomous agents research that includes biological and robotic autonomous agents. Therefore, SIA research can help with science of social minds.
- Autonomous agents definition [Franklin/Graesser]: Is a system in an environment, senses it and acts on it, over time, following its own agenda, and affecting what is senses in the future.
- SIAs are agents (of any kind) that show elements of human-style social intelligence. Artificial SI is human-style SI in artificial agents.
- Note that SIAs do not have to be intelligent!
- [Fogg] Computers as *Persuasive Technologies*: technology that is designed to change a person's attitudes or behavior in a predetermined way. *Captology* is study of planned persuasive effects. Functional triad:
  - Social Actor (creates relationships)
  - Tool (increases abilities)
  - Medium (provides experience)

- Life-like Agents Hypothesis: "Agents that should interact with humans are most successful if they imitate life. This can be (a) focusing on presentation and believability (shallow approach) or (b) model cognition and intelligence (deep architectures)
- Life-like agents are good because
(1) They act on behalf or in collaboration with the user, fill human roles, and therefore require human forms of (social) intelligence
(2) More human looking and acting agents are preferred by humans, so life-like agents could be pets or personal assistants.
(3) Life-like agents could serve as models for scientific investigation of animal behavior and minds.
- Problem with (1) and (2):
(a) Mimicking human appearance can restrict the functionality of the agent! New designs (not imitating) could be better for new functionality.
(b) Human appearance could make user expect more, like human intelligence and personality.
(c) Hard to integrate new functionality in a real-looking agent plausibly, without making it look less human! (Why can this salesman fly??)

Attitudes towards SIA: Anthropomorphisms and Behavior Reading

- Social Intelligence Hypothesis: Intelligence comes from Social Intelligence which itself was necessary to deal with complex social situation of humans. There was a transfer from social to non-social intelligence.
- Why do we anthropomorphize?
  - With animals [Eddy]:
    (1) because of physical similarity
    (2) because of existing attachment (familiarity) (dogs are more familiar than frogs)
  - Other evidence
    (3) because behavior-in-context.

Attitudes towards SIA: Robot case study

Dautenhahn's experiment with kids, how the describe and interact with robots. Results: Kids...

- tend to give robots human face (and feet..)
- tend to put robots in familiar settings
- tend to anthropomorphize robots

Societies of Social Animals

- Swarm intelligence: non-individualized societies, using stigmergy (indirect communication via the environment)
- In primate societies, we have individualized society.
  - Socially situated: being part of and surrounded by social environment
  - Socially embedded: pay attention to others and interact individually
- Humans are very good at that: predicting, manipulating, and dealing with social dynamics, involving direct and third-party relationships. Method is *social grooming*, and they communicate with *stories*. They seem to have mental models of themselves and others.
- *Social Intelligence Hypothesis*: Primate intelligence comes from adaptation to social complexity! It is actually a feedback loop, because primate intelligence leads to increased brain size that in turn enabled dealing with even more complex societies.
- *Mindreading*, *Theory of Mind*: reflect on own mental state and those of others.
- Problem with Social Intelligence Hypothesis: can account for primate intelligence, but not specific human intelligence. Solution: Stories are most efficient and natural human way of communication. Therefore:
- *Narrative Intelligence Hypothesis*: gossip, communication about third-party relationships, differentiates us from other primates. We use our mental capacities to reason about other agents and social interactions. But that might be actually also true for other species (gray parrots)

Primate culture: we are not alone

- Three systems of memory organization are human specific:
(1) Mimetic skill

(2) Language
(3) External symbols
- Leads to language, gestures, local rituals, images.
- BUT: Others say that culture is not unique to humans: e.g., sweet potato washing skill in Japanese Macaca, invented 1953, then passed down to successive generations, is a tradition.

Implications for Evolvability of Human Societies

Evolution of primate culture:
(1) Individualized societies
(2) Social networks
(2a) Direct relationships
(2b) Identifying third-party relationships
(2c) Recognition of conspecifics
(3) Mechanisms of social bonding, either with physical or social grooming
(4) Social learning: use others as 'social tools'.
- Interesting: language is 2.8 times more efficient than physical grooming for social bonding: ideal conversational group size is 3-4)
- Summary: biological evolution leads to anonymous and individualized societies. Individualized societies are prerequisite for culture, which is necessary for social learning. 150 is the amount of people we can have direct relations to (without technology), and 3.8 is the perfect group size for conversations. These numbers have to be kept in mind when designing social agents that fit human cognitive demands.

Social Robots in Rehabilitation

- Autism impairments: (1) social interaction, (2) social communication, (3) imagination.
- Autonomous robot as therapeutic tool.
- Non-humanoid robot is better because autism patients cannot read facial expressions and are confused by them, because people seem to be unpredictable.
- Hypothesis: (1) autistic child is interested in robot (2) robot can demonstrate aspects of human-human interaction (3) robot can guide children towards more complex forms of social interaction.

Mindreading

Human social competence is based on three units:
(1) **Intentional system**, identifies self-propelled movements in space
(2) **Social system**, specifies the changes that the intentional objects undergo
(3) **Theory of mind system**, outputs explanations, states of mind, etc, to explain actions

Mindreading (for kids older than 4) **is similar**:
(1) Intentionality detector (ID)
(2) Eye-direction detector (EDD), together with ID enables dyadic representations
(3) Shared-attention mechanism (SAM), enables triadic representations
(4) Theory-of-mind mechanism (TOM), visible through pretend play.

Interaction dynamics

Dynamics between caretakers and babies.

Aurora preliminary results

- Comparative trials, with robot and non-robotic toy.
- Kids more interested in robot, and interested in front which looks eye-like.
- Therefore, Aurora can build on ID and EDD.
- Robots are only one potential form of therapy...
- Problem with results: Autistic kids cannot give verbal feedback.

Empathy

- Fundamental mechanism to bond. Distinction [Wispe]:
  - Empathy = way of knowing
  - Sympathy = way of relating

[[skipped a lot of things]]

- *Social Currency* is the glue that connects two agents empathically.

Comments

[none right now]

David Stork (ed.) (1997). HAL's legacy: 2001's computer as dream and reality. Cambridge MA: The MIT Press (book, 384 pages, chapters 1,2 and 9)

July 30/31, 2001

Summary

Preface

This book is about "How realistic was HAL?" What is the movie good for?
- Analysis: watch the movie closely
- Teaching: Movie illustrates scientific disciplines
- Prognostication: What are the most promising approaches toward AI?
- Reflection: What came true of this science fiction?

Chapter 1: The Best-Informed Dream

Most science fiction movies are very wrong, if one analyses them closely. 2001 is different, made very carefully. The movie is a meditation on the evolution of intelligence. But actually why don't we have an AI yet?

- Chapter 2, Minsky: we still do not understand learning, reasoning, and creativity.
- Chapter 15, Wolfram: missing pieces are in the domain of *complex systems*: like water flow around a rock can be explained with simple laws of physics applied to simple molecules, complex cognitive systems are can be explained with simple rules of nerve cells. It is more about algorithms than equations, though.
- Chapter 7, Kurzweil: Just scan the brain and reverse engineer it!
- Chapter 9, Lenat: First teach computers common-sense knowledge, then it can keep learning on itself.
- Chapter 16, Dennett: we need a computer to explore the world, and learn from its interactions.

They all agree that it is possible in principle to create an AI. In some domains, we have already surpassed HAL:

- Chapter 3, Kuck: According to Moore's law we soon can build HAL.
- Chapter 4, Iyer: Fault-tolerant computers can be built, especially the hardware.
- Chapter 5, Campbell: Analyzes the chess scene.

- Chapter 7, Kurzweil: Applies his phonological speech-reco system to the original soundtrack. It works with only two speakers and silent a spaceship, but in general needs more semantics, common sense, context, and world knowledge.
- Chapter 11, Stork: Speech-reco combined with lip-reading--but still far from successful lip-reading.
- Chapter 10, Rosenfeld: Visual pattern reco is far from producing verbal description of sunset!
- Chapter 6, Olive: Natural sounding speech generation seems to be harder than expected. Artificial speaking requires content understanding.
- Chapter 8, Schank: Language understanding can be faked easily (Wizard of Oz, see ELIZA), but we are still far from real general intelligence.
- Chapter 14, Wilkins: Planning is hard! HAL was not very good at it (navigate ship, search for extraterrestrials, kill crew), and neither are current computers.
- Chapter 13, Picard: Emotions: would you trust affective computers?
- Chapter 12, Norman: Human and machine intelligence has to include emotion, not only logic: softer cognition, related to emotions and making mistakes.

Other issues:
- HAL has more personality and emotions than the astronauts. But can it be punished for murder, without trial? Dennett: moral responsibility requires higher-order intentionality, and HAL shows that.
- Shall we make computers more intelligent by mimicking the human brain or by exploiting their particular strengths? Extreme on one end is Kurzweil, and commonsense approaches.
- Society's perception of computers: is the public not worried about them anymore?
- Why are there no PCs and not PDAs? Why are there no networks?

--> We have met some visions of HAL: speech, hardware, planning, chess; but not in domains like language understanding and common sense.

Chapter 2: Scientist on the Set: An Interview with Marvin Minsky

- AI needs several parallel knowledge representations, so that if the system is stuck it can jump to another.
- It's bad to build physical autonomous robots, when simulations would advance us much faster.
- Three approaches to AI (1) case-based (2) rule-based (3) connectionist reasoning
(1) Program has database of stored problems and solutions. When new problem comes up, it tries to find a similar problem. Very difficult to do.
(2) Program has a large number of rules. It is difficult to cover every possible input. These systems are brittle.
(3) Learning rules in big networks of simple components, inspired by brain. It is not clear how a network solves a problem, which is bad.
- 2001 is good because humans are made look so shortsighted, and HAL stole the show.
- General Intelligence: No one has tried to make a thinking machine and then teach it chess. We haven't progressed toward a truly intelligent machine; we only have some dumb specialists in small domains. One of the dumb ideas is Situated Action Theory: there are no internal representations of knowledge, just the current data from the sensors.
- Emotion: Roz is working on detection of emotion and giving computers emotional output. Minsky is more interested in reasoning about emotion. Emotion is important in switching between sub goals of an AI, when to change the approach. (HAL detects and shows more emotion than any other character in 2001.) Stork: But computers with emotions are not reliable! Minsky: The humans make the errors, eventually.
- Minsky: If we work hard and smart, we can have HAL in between four and four hundred years...

Chapter 9: Common Sense and the Mind of HAL

- AI is based on goals. Some AI generates its own goals. However, there are three goals that computers should NOT have:
(1) Kill everybody around you, extreme violence (HAL, Shakespeare, Eastwood)
(2) Do nothing, commit suicide (Eurisko)
(3) Be omniscient (God)

- HAL was stupid, because if it had been smart, it would have found another solution, just as we do every day (like lying, or trusting people).
- Important: accept inconsistency. Occasional mistakes are inevitable and can even be a learning experience. Sources of inconsistency:
  - Different levels of generality and precision (sun revolves around earth in daily life)
  - Different epistemological stati: there are no vampires; Dracula is a vampire
  - Conflicting learning experiences of different times (phobia as a child, still existing as an adult)
  - Inconsistencies passed to us from others
- *Brain as a knowledge pump*: One has to invest time in training before being able to do a task, e.g.,
  - Leading a space mission requires
  - Engineering degree, requires
  - Undergrad school, requires
  - High school, requires
  - Elementary school, requires
  - Experiences as baby and toddler: e.g., food is usually prepared in kitchens; chairs are for sitting, etc.
- Before one can learn to talk, and survive on one's own in the everyday world, one needs "common sense."
- We need common sense to:
  (1) resolve ambiguities ("Can you hold it closer?")
  (2) deal with (human) time ("Is the radio still dead?")
  (3) metaphors
  (4) colloquialisms
- We do not only need common sense, but we have to assume that the other party does as well: we need "shared understanding." That enables us to use very short sentences, which is even more extreme between twins, married couples, etc. Computers to not have the context of human life.
- Common sense is used not only in language, but also in writing and action.
- Common sense helps to restrict reasoning space (excluding irrelevant data) ("Where is my Visa card?")
- Common sense also "accumulates" like a snowball.
- Common sense is the foundation of "consensus reality."
- If HAL were to be shut down, his ability to hold a conversation would have been the first one to leave, not the last one.

- How to build HAL in three easy steps:
  1. Prime the pump with Common Sense knowledge: millions of everyday terms, concepts, data, rules of thumb, etc.
  2. On top of that, construct the ability to communicate in English, and let it enlarge its knowledge base like that.
  3. Once it knows "everything" (has talked to everybody), it will have to make experiments on its own. ((Does this mean: Give it physical embodiment?? Or give it power to change things, like the executive power of a company or country??))
- These steps are additive, not sequentially exclusive.
- Problems with Step 1:
  - What is this data to be entered?
  - How should it be represented in the machine to allow efficient deductions?
  - Who will enter that data?
  - If many people, what will keep them from contradicting each other?

CYC: Doug Lenat's experience with Step 1 over the last past dozen years

- Started with step 1. End of 90s should come step 2, at 2001 step 3
- Goals was NOT to understand how human mind works, and NOT to test some theory of intelligence, but just to build an artifact ((that works))

What should CYC know?

- NOT encyclopedia directly, because that's the complement of CS knowledge!
- BUT: Explain each sentence in an encyclopedia.
- AND: Explain the leaps between two sentences of an encyclopedia entry, what reader infers between the sentences
- Until 1996, they used such bottom-up approaches. Then they started top-down, telling CYC about topics.

How should knowledge be represented in CYC?

- Earlier: frame-and-slot language: "TimeOfBirth (HAL) = 1/12/1992"
- Later: second-order predicate calculus.

Lessons learned

(1) **No probabilistic weights**. CYC can gather all the pro and con arguments, examine them, and then come to a conclusion
(2) Every representation is a **trade-off** between:
- Expressiveness: how easily you can say complicated things, e.g., English; epistemological problem
- Efficiency: how easily the machine can reason with what you told it, e.g., programming languages; heuristic problem

Solution: Separate these two! They created two languages:
- EL (epistemological language) for the knowledge enterers. This is then converted to
- HL (heuristic language) for efficient reasoning about time, causality, containment, etc.

(3) **Don't try to be consistent in the whole knowledge!** Solution: Separate knowledge into hundreds of contexts and micro theories. Within these, the knowledge is consistent, but there might be contradictions among them. E.g., jump and scream when something good happens: appropriate in Football game context, not appropriate in office context. Different eras are: time, political or religious points of view, etc.

Applications of CYC

- **Knowledge fusion and integration.**
  CYC will have to be able to use semi-automatic knowledge acquisition, e.g., from databases and tables. One application would be to do a "sanity check" on this information, e.g., crosscheck information from several databases (X can't be his/her own spouse, X can't be suspected of a crime and be in jail at the same time). CYC would treat each column separately, writing rule that explain its meaning.
- **Information retrieval.**
  Image content search, without previously attributed keywords. (Today, software could keyword searches in caption, using thesauri for synonyms and dictionaries for definitions.)
- **Natural language understanding.**
  Web search machines that understand the web content. CYC knows that "a happy person" could be the same as "a man watching his daughter take her first step."

Conclusions

- Essence of common sense: "A little goes a long way."
- One can't bead Gary Kasparov in chess with common sense, but with the right common sense knowledge, one can make most of the daily inferences easily; if one lacks it, one can't solve these problems at all. Ever.

Comments

[none right now]

Kerstin Dautenhahn (1999). Embodiment and Interaction in Socially Intelligent Life-Like Agents (book chapter, 40 pages)

August 3, 2001

SKIPPED

Anne Foerst (1995). The Courage to Doubt: How to Build Android Robots as a Theologian (talk, 7 pages)

August 3, 2001

Rapid Summary

- Describes COG, as an example for embodied AI research.
- Opponents say that the soul is something completely different from the parts of the bodies; it does not emerge from the interaction of the different parts.
- How to make those two incompatible camps talk to each other?
- Paul Tillich's Theory of Sin and the Courage to Doubt.
- Human life is a balance act between
  - Individualism vs. participation
  - Dynamics vs. form
  - Freedom vs. fate
- Theologians can accept the worldview of AI (everything can be described mechanistically), but they know that it can't be proven right or wrong, like the relationship between God and humans.
- Outcome: Let's combine theology and AI science! One cannot explain how the brain works, and the other cannot give any answer about meaning of life etc.

Comments

I don't think there is much in this talk that I can use. Perhaps there are better papers?

Robert D. Putnam (1995). Bowling Alone: America's Declining Social Capital (paper, 13 pages)

August 3, 2001

Summary

- Social capital = features of social organizations such as networks, norms, and social trust that facilitate coordination and cooperation for mutual benefit.
- Social capital in the US (and in other countries) has eroded over the last generation.
- Parts of this essay:
  (a) What's good about social capital?
  (a) How can we tell that it is decreasing?
  (b) Why is this happening?
  (c) What can we do about it?

What's good about social capital?

- most fundamental form: family; also: neighborliness
- produces better schools, faster economic development, lower crime, more effective government
- "life is easier in a community blessed with social capital"
- facilitate coordination and communication
- the individual's sense of self is broadened, from 'I' to 'we'.

How can we tell that it is decreasing?

- Fewer voters, less engagement in politics, people are less trusting
- Fewer members in
  - church-related groups
  - school-service groups (parent-teacher associations)
  - labor unions
  - Boy Scouts and Red Cross
  - bowling leagues (but more Americans are bowling!! bad for business, because members of leagues consume three times more pizza and beer than solo bowlers)
- Countertrends, where there are more people:
  - *Tertiary associations*: environmental organizations, feminist groups, AARP (seniors); they have big political influence, but members do not really participate other than paying annual fee.
  - *Nonprofit organizations*
  - *Support groups*: not fostering community, because "come if you have time, talk if you feel like it, respect everyone's opinion, never criticize, leaf quietly if you become dissatisfied..."

Why is this happening?

- Movement of women into the labor force: lead to decline in women's civic participation.
- Mobility: the 're-potting' hypothesis. We have more cars and are more mobile. Problem: we also have more homes.
- Other demographic transformations: fewer marriages, more divorces, and fewer children
- Technological transformation of leisure: TV! It makes our communities wider and shallower.

What can we do about it? Above all: more research!

- What are the dimensions of social capital (it's NOT one-dimensional)?
- What types of organizations and networks most effectively embody or generate social capital?
- Macro sociological crosscurrents: effects of electronic networks on social capital?
- We must not romanticize the 50s! There are downsides to social capital, e.g. more corruption
- Impact of public policy? E.g.,
  - slum-clearance has destroyed a lot of social capital
  - consolidation of country post offices and small school districts
  - Not to forget the POSITIVE public policy impacts: community colleges, tax deductions for charitable contributions, law that 'all houses have to have front porches.'
- What is happening in other countries? Is comparable erosion under way in other advanced democracies?

Comments

I am not sure if this paper is relevant for me, I can't really connect it back to the rest. I will probably skip the book for now.

Douglas R. Hofstaedter and Daniel C. Dennett (1981). The Mind's I: Fantasies and Reflections on Self and Soul

August 8, 2001.

Summary

Introduction:

- Philosophical issues about teleportation: Is it a murdering twinmaker?
- Me vs. my body: If I *have* a body, so I must be something *other than* this body.
- Me vs. my brain: If I *have* a brain, so I must be something *other than* this brain.
- Does thinking happen between my eyes, or is it just because I have this optical perspective?
- Am I more than my body/brain?
- Is my soul a ghost, something mystical? But there are other things that are neither mystical nor ghostly, but don't are physical either and not only abstract objects: e.g., haircuts, a hole, and the game of bridge. They can have birthplaces and histories. Therefore, one doesn't have to believe in ghosts to believe that selves that have an identity which transcends the living body.
- What is consciousness? Are other animals conscious? Babies before birth? Computers or robots? Are we conscious when we dream? Can a human have more than one conscious subject (agent) in one brain?
- Perspectives: from the inside, the consciousness is clear. From the outside: not so clear. Simple with ourselves: we can observe the coincidence of one's inner life with one's outwardly observable behavior. But that's different with others, we only have the observable behavior. If a robot could tell us about its inner life, would we accept that? Is that special inner light really turned on? In general: enough of the right sort of outside facts will settle the question of whether or not some creature is conscious. Just like with a car: Is there REALLY an internal combustion engine in there?
- [Locke]: Introspection: Nothing is hidden from the inner view.
- [Freud]: There are unconscious mental processes that are hidden. These processes belong to other "selves!" (Axiom still true that every mental state must be someone's conscious mental state.)
- Cognitive psychology: information processing (sophisticated hypothesis testing, memory searching, inference) can be beneath our conscious level, inaccessible to introspection. Some of these activities are even more accessible to outsiders than the owners!
  -> Problem: Then there must be unconscious subsystems in our brain, "nonconscious bits of organic machinery." What is consciousness for if unconscious information processing can do all the stuff we though consciousness is necessary? Therefore, robots could do that too! How could any subjectless information processing add up to what is so special with humans? What's the difference between unconscious info processing and conscious thought itself, directly accessible? (Unconscious processes can be accessible to other processes too!)
- Is it possible that we have several consciousnesses in one brain? -> Split-brain cases: they have two independent minds, but are both dominant and non-dominant halves conscious.
- Research says that even unattended signals are processes (channel comprehension): does this mean we have at least two different and only partially communicating consciousnesses?

Chapter 4: Computing Machinery and Intelligence (Turing)

- Main question: Can machines think? He replaces the question with Imitation Game. (Man, woman, and interrogator who has to find out who has what gender. Only one person helps the interrogator.) Digital computer could replace one of the players.
- Nine objections:
(1) Theological Objection: Thinking is a function of man's immortal soul. No animal or machine has a soul, and therefore can't think. Isn't this a restriction of the omnipotence of the almighty God, which can't be?
(2) Heads in Sand Objection: Consequences of a thinking machine would be dreadful, let's hope it will never happen.

(3) Mathematical Objection: Goedel's theorem: In any sufficiently powerful logical systems statements can be formulated that can be neither proved nor disproved within the system. And this would not apply to humans. (But there is no proof for THAT.)
(4) Argument from Consciousness: No mechanism could feel pleasure at its success. This argument denies the Turing test. Perhaps use the simplified test, with only two players: one tries to figure out if the other REALLY understood something of the conversation, *viva voce*.
(5) Arguments form Various Disabilities: "The machine can certainly do X, but not Y." Y = be kind, resourceful, beautiful, friendly, have initiative, enjoy strawberries and cream, etc. Special case: Machines cannot make mistakes. Of course they can: errors of functioning, and even errors of conclusion (weather forecasts)
(6) Lady Lovelace's Objection: A computer cannot never do anything really new. Not true: machines can surprise, and they do often.
(7) Argument from Continuity in the Nervous System
(8) Argument form Informality of Behavior: It is not possible to produce a ser of rules that describe what a man should do in every conceivable set of circumstances. This does not happen to real people, so they can't be machines.
(9) Argument from Extrasensory Perception: Machines do not have ESP like telepathy, telekinesis.

Reflections on Chapter 4:

Turing thinks ESP is the strongest argument, but ESP does not exist according to today's psychology. If they would exist, it would cause a major revolution in the scientific world, the current laws of physics cannot be amended

Comments on Chapter 4

Of course there are ways to integrate ESP in current scientific world, just look at quantum physics and other unusual scientific domains.

Chapter 5: The Turing Test, a coffeehouse conversation (Hofstaedter)

- Three people: physics student, biology student, and philosophy student talk about Turing test.
- Many people are upset by the suggestion that people are machines, or that machines might think.
- Turing test is like IQ test for machines? How do we know that the Imitation Game will get at the essence of thinking?
- If man wins Imitation Game, then he has good insights into feminine mentality. If computer wins the game, it has good insight into what it's like to be human. But that's not necessarily the same as thinking! It could be simulation of thinking!
- Is a simulated hurricane a hurricane? But does a cash register really calculate? Do humans calculate, or just manipulate mental symbols? (Dennett's reductio ad absurdum.)
- Thinking can take place in an of several billion brains, which are physically different. Important is the *pattern*, not the medium. So we can come up with a new medium, and thinking could still happen. So it is about internal structure, organization (not external manifestations)
- Another approach: test emotional responses: can it understand a joke? Thought and emotions can't be divorced. The ability to think, to feel, and consciousness are just different facets of one phenomenon. But do certain animals show emotions (dogs, cats).
- Other idea: consciousness requires a certain way of mirroring the external world internally, and the ability to respond on the basis of the internal representation. And it has to have a self-model.
- Intelligence requires motivations. BUT: When one puts enough feelingless calculations together in a huge coordinated organization, then one gets something that has properties on another level, a system of tendencies and desires and beliefs. One has to switch the level of description, which Dennett calls "adopting to the intentional stance." But it gets interesting if computers itself adopts the intentional stance towards itself.
- Can one have the intentional stance towards being other than humans? Mammals? But there can be thinking being that would fail the Turing test.
- But humans have and "internal flame", are creative!
- Current computers do not think, but could future computers?
- Making mistakes is a sign of high intelligence.
- Missiles that decide that they are pacifists; smart bullets that don't want to commit suicide...
- Parry: program that has already passed a rudimentary version of the Turing test. But it acts paranoid and wants to maintain control. Selects from a huge repertoire of canned sentences the one that best responds to the input sentence.

Reflections on Chapter 5

Turing test is based on behaviorism, operationalism, and verificationism. It tests if one can act/behave intelligently. It seems that now one might build a machine specifically for the Turing test, but it can do nothing else. Is the Turing test bad because of its *black box* ideology? No, because we treat each other also as black boxes, and this ideology is used in all ((?)) scientific investigation.
Another problem is representation: A representation of a cow is not a cow, but is a representation of a mathematical proof not a mathematical proof? (Even a cartoon phony proof might simulate something about the mathematician: verbal mannerism, absentmindedness, etc.) But where falls the mind: is mentality like milk or like a song?

Comments on Chapter 5

The style of this chapter is rather unstructured--nice to read, but bad to get structured knowledge out of it.

Chapter 7: The Soul of Martha, a Beast (Miedaner)

- This is about a chimpanzee, Martha, which can talk via a direct neural interface.
- She turns out to be a very believing, trusting, happy, child like character.
- It is intelligent in a broad sense, even in a human sense. Martha is more intelligent than a human idiot and human imbecile.
- Human like intelligence does not correspond to human like treatment.
- After such lab animals have outlived their usefulness, they get eliminated.
- The researcher demonstrates that in court, and Martha gets killed by a poisoned candy, but during her death her brain expresses her pain (Hurt Martha Hurt Martha), and then astonishment (Why Why Why), which is absolutely heartbreaking.

Reflections on Chapter 7

What is the difference between having a mind (intellect) and having a soul (emotionality)? Can one exist without the other? Is the degree of intellect a true indicator of degree of soul? Do retarded or senile people have "smaller souls" than normal people? Can we measure the soul through language? Is the Touring test a soul meter?

Comments on Chapter 7

Very interesting approach to mind, over emotionality. I think most people would agree on the argument that a soul is strongly linked to emotions, but that transfers the problem only to the question of how to detect true emotions, vs. simulated emotions.

Chapter 8: The Soul of the Mark III Beast (Miedaner)

- Assumption: Biological life as a complex form of machinery. Therefore, machines are just another life form.
- Illustration: Can one relate to a machine? Is it possible to be bothered by breaking a machine, like killing an animal?
- A person who cannot kill an animal is not necessarily against killing, but rather bothered by doing it herself.
- Killing an animal is hard because it resists death: it cries, struggles, or looks sad. Therefore, it's the person's mind that has a problem with killing.
- Illustration: Trying to kill an animal like machine is very hard, if the mechanical animal exhibits animal like behavior: sucking current from outlet = eating, obstacle avoidance = evasive behavior, leaking lubricating fluid = bleeding. From the human emotional perspective, destruction of such a machine is not different from killing an animal.

Reflections on Chapter 8

Humans can assume without problems that there are mechanical, metallic feelings. The question if one can kill an animal depends also a lot on the circumstances (drowning and an in the sink, feeding live food to reptiles). We all sense that there is soul-killing going on in slaughterhouses, but we don't want to be reminded of it. Humans are all animists to

some degree, but the souls we project into these objects are an image purely in our minds. We have a "storehouse" of empathy that we can tap into more or less easily; fleeting expressions, etc can soften us.
When does a body contain a soul? Not as a function of the inter state, but as a function of our own ability to project. This is very behaviorist, since we ask nothing about the internal mechanisms. This is a strange kind of validation of the Turing test as a "soul detector".

Comments on Chapter 8

This very short chapter is probably more valuable than most of the articles that I have read if it comes to assessing how people might react to autonomous beings: Basically, if there are enough (small) cues so that we can project a soul into this entity, it will HAVE a soul, which in turn can raise it to a completely accepted being with all human privileges. The question is only: how few cues do we actually need? Probably even less than what Nass and Steuer suggest, if the context is appropriate.

Chapter 10: Selfish Genes and Selfish Memes (Dawkins)

- Modifying Darwin's "Survival of the Fittest" to "Survival of the Stable"
- Stable patterns of atoms lead to molecules. This doesn't need any magic, it just happens in a weak brown soup of water, carbon dioxide, methane, and ammonia, plus UV light.
- This lead to larger molecules, and eventually to a Replicator molecule, which seemed to be a very unlikely accident. True, very unlikely, but it had to happen only once, because after that, it would create more and more replicator molecules from this template.
- Because there were errors in the replication process, variations occurred. They were different in the following ways:
  (a) Longevity: more stable, live longer
  (b) Fecundity: speed of replication
  (c) Copying-fidelity: accuracy of replication (Note that although evolution seems to be a "good thing," nothing actually "wants" to evolve, it is just happening.)
- So are these molecules "living"? We have no answer, and it doesn't make any difference anyways.
- Because the primeval soup is eventually limited, it can support only a limited amount of molecules, so there was *competition*: the molecules tried to increase their stability and decrease the stability of the competitors, e.g., by breaking up molecules chemically (proto-carnivores).
- An important factor in this competition was the development of *survival machines* for replicator molecules to live in: protective coats, etc. These survival machines were getting better and better, developing muscles and hearts, eyes.
- And here we are: we are the most perfect survival machines for our replicator molecules, the genes.
- Modern replicators live in packs (clans), being actually rather a *gene complex*, and they have developed a hierarchy where some genes control other genes.
- Sexual reproduction deals with the multiplication of survival machines, but also shuffles genes into different gene complexes. A certain gene combination may be short-lived, but the genes themselves are potentially very long-lived. Genes travel from person to person, from generation to generation, but do not always influence the survival machine itself. But a gene does not get senile. They are somehow immortal: they live thousands of millions of years.
- Survival machines were first just walls, and they fed on freely available molecules in the primeval soup. Later, some of them started using sunlight directly (plants) or eat other molecules (animals)
- Although our bodies are colonies of genes, they have also developed individuality on their own.
- Bodies have also developed a *purposiveness*, trying to maintain an optimum state (reducing discrepancy between current state and "desired" state; they have to solve problems like oscillation, overshooting, and time-lags). This leads to life-like behavior, like a guided missile.
- Like a computer programmer is not a puppeteer of the running program (she just gives a list of specific knowledge and hints about strategies), genes too control the behavior of their survival machines only indirectly. They can't directly, because they would be too slow; that's why we have brains. (Example of Andromedans who sent us the blueprint of a computer that would control the earth.)
- Genes therefore have to predict somehow what can happen to their survival machines. One possibility to predict is to use simulations. Survival machines have developed it: Imagination. They are a step ahead of the survival machines that only learn on the basis of trial and error.
- It is possible that the capacity to simulate lead to subjective consciousness: perhaps consciousness arises when the brain's simulation of the world becomes so complete that it must include a model of itself?

- Consciousness is the culmination of an evolutionary trend towards emancipation of survival machines from their masters, the genes. The body can now even rebel against the genes. Although genes are the primary policy-makers, and brains are the executives, brains took over more and more of the actual policy decisions using tricks like learning and simulation. Eventually, the genes will probably give complete freedom to the survival machines, with the single order: do whatever it takes to keeps us alive.
- On alien planets, we would probably find other kinds of replicators. But even on our planet, a new kind of replicator has emerged. The new soup is the soup of human culture. The replicators we could call Memes. Examples are: tunes, ideas, catch phrases, cloth fashions, ways of making pots, etc. Genes leap from body to body via sperms; memes leap from brain to brain via imitation. They should be regarded as living structures. They parasitize a brain, turning it into a vehicle for the meme's propagation.
- When we die we can leave two things behind: genes and memes. The genes might be immortal, but not the specific collection we had. Therefore, it might be better to seek immortality by leaving behind meme-complexes. Socrates' genes might or might not be alive today, but his meme-complex is still very clear.

Reflections on Chapter 10

This is an extreme case of a reductionist approach: mind comes from molecules. It is based on the fact that people underestimate the complexity that can result from a lot of small units interacting very fast according to formal rules. Important is that there might be other media that support life-like or thought-like activity.

Comments on Chapter 10

It sounds very convincing, but there are a few points in his chain of reasoning that are weak: I think I can imagine a computer that simulates the external world, including its own body, and still has no consciousness. I think simulating its own behavior is an important step towards a mind, but there is an emotional component that I am missing. But perhaps it is only necessary to step up one or several more meta levels!?

Chapter 11: Prelude... Ant Fugue (Hofstaedter)

- Fermat and Fourmi's last theorems
- Record with Bach playing fugues on harpsichord, reconstructed from the air molecule movements years later.
- How to listen to a fugue? To its parts or to the whole?
- Illustrations in the score: Reductionism vs. holism vs. MU, on different levels.
- How can an anthill (Aunt Hillary) talk? Not out lout, but in writing.
- How an anthill works: individual ants seem to walk around randomly, sometimes forming teams. These teams stick together if they are big enough, and if there is something to do for them. These teams are signals that move along (and can even cross each other); there are scents along the way that provide information about local matters of urgency, like nest building or nursing. If it can contribute, the signal disintegrates into single ants that do the job. Otherwise, the team keeps going. The signal has no sense of purpose. The distribution of different kinds of ants (caste) in the anthill is crucial. The caste distribution, updated dynamically, reflects the outer world. Certain combinations of teams have the meaning of "symbols" (but they are active symbols, not like letters), and there are several levels of symbols. Any system that has mastery of language has essentially the same underlying sets of levels. Caste distribution and brain state are comparable, both are "states".
- The ants are not the most important feature of an anthill, somehow analog to the letters of a book. There is no natural mapping of an individual letter to the real world! The natural mapping occurs on a higher level. The letters are the medium, not the message.
- So, Aunt Hillary thinks through the manipulation of symbols composed of teams composed of lower-level teams composed of ants. But there is no distinct agent; the full system is the agent.
- Ant colonies are not different from brains. We also don't have access to our lower levels.
- Aunt Hillary is actually an anthill that was rearranged out of the leftovers of another anthill (after a terrible flooding), and still, it is very different: there are several distinct ways how to form a sum of parts.

Reflections on Chapter 11

This is about the soul. The participants disagree on many things, except that the collective behavior of a system of individuals can have many surprising properties. Humans have also the urge to personify organizations. But does, e.g.,

a country have thoughts or beliefs? It depends on if it has a "symbol level", or is a "representational system", which means it is:

- active
- self-updating collection of structures
- organized to mirror the world as it evolves
- built on categories
- able to keep going even if cut off from the reality it is reflecting

A painting and a mirror are not, but a program that can look at a scene and tell what elements there are, what caused the scene and what probably will follow, is.
Interestingly, the elements of such a system do not see the whole picture, like falling dominos don't know what their falling means, and keys of a piano don't know which song they are playing. It is all a question of point of view (e.g., people in a fun park rotating ring that throw a ball towards the center of the ring will experience strange forms of gravity, subjectively speaking).

Hard scientist: reductionism (space) + predictionism (time) = mechanism
Soft scientist: holism (space) + goalism (time) = soulism

Finally: Self-awareness comes from the systems responses to both external and internal stimuli. "Mind is a pattern perceived by a mind" (circular, but neither vicious nor paradoxical!) The brain needs these different levels to stay flexible.

Comments on Chapter 11

Poetic and polite, but also longish conversation between a tortoise, a crab, and an anteater, that try to explain a reductionist approach to soul. Some of the core points that are discussed are demonstrated at the same time with the style of the conversation! Neat. Most important point is probably that complex systems must have different levels of abstraction (sometimes semiautonomous subsystems), and only the topmost level could be called conscious or the consciousness. Nevertheless, the elements that create the levels underneath this top level are not "aware" of the big picture, but still essential. The chapter demonstrates such different levels also visually: several levels of writing, saying different things depending on how closely one looks, and mentions the same effect on musical level (fugues).

Chapter 13: Where Am I? (Dennett)

- Brain of a person gets taken out of the skull and placed in a vat, connected with the rest of the body with a wireless connection. Where is this person? Where the body is or where the brain is?
- After an accident, person gets a new body. No problem.
- Person gets a spare brain, a duplicate that runs as software in a computer. Can switch back and forth between computer and brain, and there is no difference, because both are connected to the body in parallel, and function exactly the same. What if there would be two bodies, one controlled by the brain, one by the computer? Which one is the original person?
- Problem: If the two brains drift apart, they will become different personalities!

Reflections on Chapter 13

This story has a lot of problems because it claims that a bio brain and a silicon brain can perform the exact same functions, independently, without synchronization of any kind. That's unrealistic.

Comments on Chapter 13

There are many technical issues that will never be solved, I guess, but still, separating body and brain physically is an interesting thought experiment.

Chapter 18: How Trurl's Own Perfection Led to No Good (Lem)

- Trurl finds a king without kingdom on an asteroid, and creates one for him in a box. It is perfect (because he is Trurl, a constructor), and can be manipulated with knobs from the outside. The king takes it and rules as a tyrann over this micro world.
- Back home, Trurl's friend tells him that he is responsible for the entire bad thing the brutal despot and slave driver does to the miniature world, and that Trurl has committed a terrible crime. A sufferer is one who behaves like a sufferer!
- Trurl wants to go back and hold an election. But the civilization has already broken out of the box, populated the whole asteroid, and got rid of the King: made him to a moon of their world!

Reflections on Chapter 18

Compare this little world with Tom Robbins miniature Tyrolean village, having all the details, e.g., an orphanage that burns down every Xmas eve. However, it's not the same because the repetition of the drama robs the little world of any real meaning.

Comments on Chapter 18

Weird story (written in Petit Prince style?) about an omnipotent person who could create a perfect model of reality in a small box, and what it means.

Chapter 22: Minds, Brains, and Programs (Searle)

- First makes distinction between strong and weak AI: former says that computers can understand and have cognitive states; latter says that computers can be very useful tools to study the mind (but do not have one).
- Criticizes Schank's program that claims to understand stories (or rather: strong AI says that it understands stories, weak AI says that it can explain the human ability to understand stories)
- Chinese room counterexample: isolated person manipulates stacks of symbols according to (English) rules; to the outsider, it looks like the person answers questions asked in Chinese.
- Compares person who answers English questions, and participates in Chinese Gedankenexperiment. In the first case, the person really understands the story and can explain answers. In the second case, the person neither understands the story nor can explain the answers.

Reflections on Chapter 22

None

Comments on Chapter 22

IMHO not very well written article: the author repeats his arguments over and over, but they do not get better through repetition. I also think one has to look up what Schank's program actually does in order to understand Searle's objections and especially the nine counter objections.

Byron Reeves and Clifford Nass (1996). The Media Equation. (book, 317 pages, selected chapters)

August xx, 2001.

Summary

Comments

I think I am aware of the main points of this book.

Joseph Weizenbaum (1976). Computer power and human reason: From judgment to calculation. (book, 300 pages, selected chapters)

August xx, 2001.

Summary

Comments

Joseph Weizenbaum (1966). ELIZA: A Computer Program for the Study of Natural Language Communication Between Man and Machine. (paper, 10 pages)

August 30, 2001.

Summary

- How it works: Input sentences are analyzed for keywords, and then decomposed according to rules. Responses are generated by reassembly rules associated with decomposing rules.
- Fundamental technical problems:
    - Identification of keywords
    - Discovery of minimal context
    - Appropriate transformation
    - Generation of responses without having keywords

- ELIZA acts like a Rogerian psychotherapist. It works only in this specific environment of psychiatric interview because one of the parties is allowed to know almost nothing of the real world. (E.g., user says "I went for a long boat ride," the doctor can ask "Tell me about boats" without being looked at like an idiot.) Note that the user must make this assumption, otherwise, it will not be credible.
- Credibility is important: How can credibility maintained over a long time? Can one decrease credibility gradually to find out what is the minimum level required? Does the initial instruction play a role?
- ELIZA is elegant because the illusion of understanding comes from very little "machinery".

- Future plans:
    - Give ELIZA information about the real world ((Common sense?))
    - Give ELIZA the power to build user models ((To get personalized to a user?))
    - Give ELIZA the ability to make inferences. If user says that he is not married, and later speaks of his wife, the he must be divorced or widowed.)

--> "To explain is to explain away": If something can be explained, it disappears. If there is a wondrous machine, and one suddenly can explain it inner workings (in language sufficiently plain do induce understanding), then its magic crumbles away, revealed as a mere collection of procedures, each quite comprehensible.

Comments

What happened to the future versions of Eliza? Why are there no more more advanced Eliza's around? Julia probably would be a grand-grand-grand daughter...

Here is some online versions:
http://www-ai.ijs.si/eliza-cgi-bin/eliza_script
http://www.parnasse.com/cgi-bin/drwww.cgi
http://pandi.20m.com/games/elizav2.html

Daniel C. Dennett (1987). The Intentional Stance (book)

August xx, 2001.

Summary

"An intentional stance refers to the treating of a system as if it has intentions, irrespective of whether it does. By treating a system as if it is a rational agent one is able to predict the system's behavior. First, one ascribes beliefs to the system as those the system ought to have given its abilities, history and context. Then one attributes desires to the system as those the system ought to have given its survival needs and means of fulfilling them. One can then predict the system's behavior as that a rational system would undertake to further its goals given its beliefs. Dennett argues for three main reasons for taking an intentional stance. First, it fits well with our understandings of the processes of natural selection and evolution in complex environments. Second, it has been shown to be an accurate method of predicting behavior. Third, it is consistent with our folk psychology of behavior."
http://web.psych.ualberta.ca/~mike/Pearl_Street/Dictionary/contents/I/intentional_stance.html

Another, longer summary:
http://cognet.mit.edu/MITECS/Entry/dennett

Comments

See the many references in these summaries to this book.

Bill Joy (2000). Why The Future Doesn't Need Us (article)

August 30, 2001.

Summary

- New technologies, robotics, genetic engineering, and nanotech, will replace our species.
    -> More precisely, the consequences of enormous computing power will destroy us.
        -> Even more precisely, the consequences of our truth-seeking behavior will destroy us.

- [Theodore Kaczynski, the Unabomber] Two options:
(I) Machines will be in control. The human race might easily permit itself to drift into a position of dependence on the machines, without having the choice to ignore the machines decisions.
(II) Humans stay in control.
  - Tiny elite might exterminate the (useless) mass.
  - Tiny elite might reduce the birth rate (with psychological or biological means) so that the mass goes extinct
  - Tiny elite plays the role of good shepherds. Humans are happy, but not free.
- [Hans Moravec]: In a completely free marketplace, robots (a new robot species) would outperform humans, and we would be squeezed out of existence.
- [Danny Hillis] All true, but the changes will come gradually and we will get used to them. Problem: we will loose our humanity in the process.

Dangers:

- **Destructive self-replication**: The dangerous thing about robotics, genetic engineering, and nanotech is that they can self-replicate. Genetically engineered life forms, or nanotechnologically created life forms (from "assemblers", Engines of Creation/Destruction) could destroy our biosphere on which all life depends.
- **Arms race** in genetics, nanotech, robotics (like with nuclear bombs in the 20th century).
- **Pursuing immortality**. We should think about it again.

Solutions:

- Move to **another planet**. (We probably will take our problems with us...)
- **Shields** against dangerous technologies.
- **Limit development of technologies** that are too dangerous (genetics, nanotech, robotics) (and avoid an arms race!). Don't pursue certain kinds of knowledge. Be aware that truth seeking can be very dangerous.
- Look for something else that makes us **happy**, other than material progress and knowledge.
- BTW, knowing these dangers is not a replacement for acting!

Comments

For my area, the most relevant part of this article is probably the insight that in the future, some humans will probably not like autonomous entities like robots, because these humans fear that those robots will put them out of existence.

Another relevant point for my area is the idea that robotics can get dangerous when they start to self-replicate. This in turn will influence the attitude of humans towards autonomous intelligent entities like robots. I guess that this problem is not imminent, there are only very few self-replicating robotic/software entities (e.g., viruses), and it looks like not a lot of research is pushing in this direction. The reason is probably that the entities we have are either not important/useful/universal enough to us so that we want to make them reproduce themselves, or we are happy with the life span of such entities and do not care that we have to invest money in a new one after it "dies." Or?

Bruce Tognazzini (1994). STARFIRE: A Vision of Future Computing (video)

August xx, 2001.

Summary

Comments

I have to ask Henry Lieberman for the video.

Erik Brynjolfsson and Michael Smith (2000). The Great Equalizer? Customer Choice Behavior at Internet Shopbots (paper, 50 pages)

August 30, 2001.

Summary

Compares user behavior with shopbots: sites that compare prizes from multiple retailers.

Comments

I have read the abstract and the conclusion, and skimmed the rest. I have a very hard time finding some aspects that might be relevant for my area.

User Interface Issues

Dennis Perzanowski, A. Schultz, E. Marsh, and W. Adams (2000). Two Ingredients for My Dinner with R2D2: Integration and Adjustable Autonomy.

August 29, 2001 (summarized; read much earlier)

Summary

- See also Natural Language and Gesture Interface paper by same author!
- Integration of multiple modes of communication is central to adjustable autonomy.
- Their multimodal interface for autonomous robots addresses these problems by
    - tracking goals
    - allowing both natural and mechanical modes of input: speech and gesture (both natural and artificial, from PDA)
- In interaction with robots, you don't want to repeat yourself. The interaction has to be on a sophisticated level, which includes independence, autonomy, and cooperation.
- Goals and motivation should be able to come from both human and machine. This would be a mixed-initiative system.
- For mixed-initiative systems, adjustable autonomy is crucial. There is no master-slave relationship!
- Extending human communication to incorporate humanoid robots: Humans communicate with each other on certain 'natural channels' (talking, pointing, gesturing), and when we interact with machines, we would like to do the same ((?)) At least, if the channels are provided and the robot looks human, we will use them. (It doesn't have to be that all the time, though.)
- [Griece]: Felicity makes communication effective: aptness and ease of expression incorporated in human communication.
- Gestures: They limit gestures to the meaningful ones by hand (omitting the superfluous, redundant, emotion bearing or intention bearing ones; and no facial expressions and head movements).
- Planning: Planning is necessary for collaborative work between multiple agents. They use natural language dialog and goals have been attained. The community of agents keeps a common list of goals (actions, command, directives), which changes dynamically. From the goals, plans are generated and prioritized.
- When humans interact, the build teams, cooperate, start to assume roles, learn each other's strengths, complement each other. How this works is a complex phenomenon, however: Robots have to be built that fit in this scheme! They have to become team members, and they will probably adjust their autonomy as needs arise and change. Since it is complex with humans, it will be hard for robots too. The authors say that all of these topics are related to adjustable autonomy. Goal is to build a system that is autonomous, which means that it knows enough about itself, the world around it, and what it has been doing, so the it can become a team player. When necessary, the robots can act completely autonomous, but if necessary, they interact closely with their team members. ((This just sounds like how a human works, or should work, e.g., as a police man.))

Comments

The authors have a very high level concept, and a rather simple current implementation. The idea is to make robots team players that can also act autonomously, if necessary, and adjust their autonomy dynamically. That will be very hard to realize, and will require a lot of intelligence and insight. On the other hand, the authors describe their multi-

modal interface to a robot (language, gesture, PDA buttons and PDA "synthetic gestures"), which keeps a history list of all the goals, in order to see if it already behaves like their high level goals. I doubt that, so there is still a long way to go. I like their high level goal, but it is just not realistic for now, and they don't have good ideas how to get there soon, I think.

Rino Falcone and Cristiano Castelfranchi (2000). Levels of Delegation and Levels of Adoption as the basis for Adjustable Autonomy (paper, 12 pages)

August 29, 2001 (summarized; read much earlier)

Summary

- The authors want to explore how to adjust autonomy, and they propose an explicit *theory of delegation*.
- Delegation can be based on two kinds of dependence:
    - Weak dependence: possible, but not necessary to delegate
    - Strong: delegation and autonomy are necessary, because client has no local knowledge, expertise, reaction time, or physical skills ((like memes vs. brain?))
- If an agent goes beyond the "literally" delegated task, it "over-helps". This requires reasoning about goals, plans, and interests of client. However, there is a trade-off: the more intelligent and autonomous the agent, the less passively "obedient" it is. This can lead to conflicts.

- Different kinds of delegation (A is client, B contractor), resulting in a 3-dimensional autonomy space:

First dimension (Interaction-based kinds of delegation)
- Weak delegation: A waits until B happens to realize something that fits into A's plan; exploitation.
- Strict delegation: Explicit agreement between A and B; contract.

Second dimension (Specification-based kinds of delegation)
- Close delegation: A specifies how the task has to be done before delegating it to B.
- Open delegation: A does delegate a task to B, but B figures out how to solve it.

Third dimension (delegation of control)
- Delegation of control: Full to no control: A keeps it, A gives it to B, A gives it to C, nobody controls.

Other kinds of delegation ((which don't seem to fit in the model?!))
- Conditional delegation: A says: in case of event x, B should to a task.
- Partial delegation: B can ignore the other parts of the plan of A.

- Special case: Truly distributed plan: in a combination of open and partial delegation, neither A nor B know the full plan (A doesn't care how B executes the part of the task it got, and B doesn't care about the 'big picture'): this is a *truly distributed plan*.

- Different kinds of help:
    - Literal help: B does what A told it.
    - Overhelp: B does more than what A asked for.
    - Critical help: B does what A asked for, but modifies the plan.
    - Critical overhelp: B does more than what A asked for and modifies the plan.
    - Hyper-critical help: B does not what A asked for, and changes the plan.

- Authors describe what happens if delegation does not match adoption.
- Both parties A and B can adjust their autonomy, as well as the one of the other party.

- In summary, there are two groups of autonomy:
    (1) Autonomy related to self-sufficiency: not being dependent on others for our own goals.
    (2) Autonomy related to levels of actions and goals:
        - Performance/executive autonomy: close delegation
        - Planning autonomy: open delegation
        - Goal autonomy: B can find its own goals; this kind of autonomy is not treated in this paper.
        ((Isn't there one missing? Control autonomy: less control))

Therefore, autonomy increases with
1. **more open delegation**
2. **more control delegated to B** ((this seems to me not a linear dimension, but multidimensional in itself?))
3. **more delegated decisions** (discretion) ((is this the delegation of goals, that they don't address?))

- For future work ((actually, it's in the conclusion)), they propose to merge the *theory of levels of delegation* with the *theory of the degrees of trust*.

Comments

Interesting, because it is a true theory of delegation, with an autonomy model with three dimensions, but somewhat limited, because the authors exclude the option that and agent can have its own goals to pursue. They only focus on delegated tasks and plans (although they have a section about delegation of initiative, which is not clear in itself.)

The proposed model of autonomy has some problems: the dimensions are not very distinct (might not be orthogonal), and one dimension seems to me multidimensional in itself. And Goal Autonomy does not appear at all in the model. It would be interesting to see how they combine the theories of delegation and trust.

Nice graphics, makes it a bit easier to understand, but tons of typos, and bad English, especially at the end.

Michael Mogensen (2001). Dependent Autonomy and Transparent Automatons? (book chapter, 17 pages)

August 16, 2001.

Summary

Abstract: This paper shows three perspectives on agents/autonomous agents:
1. **Construction approach**
2. **Classification approach**
3. **Use approach**

Discussion of understanding of agents and their autonomy:
(1) Designers' and users' experiences of autonomy
(2) System properties and experiences of autonomy
(3) Approaches to artifacts and action: (I) entity-oriented (II) relation-oriented

Most generic definition of agent = "Something that does something or causes something to happen"

Three perspectives on agents

(A) Design or Construction perspective [Maes 1994]

- Goes against the AI approach of "Deliberate Thinking" paradigm, which means explicit knowledge, rational choice, problem solving.
- Agents are agents because of their specific *architecture*:
  - Distributed *modules* (behaviors),
  - that work in *parallel*,
  - using *raw sensory data*,
  - without global internal model or planning activity, or hierarchical goal structure
  - Functionality and behavior are emergent properties of the agent's interaction with the environment and its modules
  - Behavior-based (not knowledge-based)
  - Bottom-up (not top-down)

(B) Classification perspective [Jennings 1998]

- Taxonomic, bird-eye view
- Agents are agents (actually, "intelligent agents") because of four attributes:

(1) Autonomous:
- operates without intervention of humans or others
- control of own actions and internal states

(2) Responsive
- perceive the changing environment and respond in a timely fashion

(3) Proactive
- take initiative
- goal directed behavior

(4) Social
- interact with other agents and people
- help others with their activities

More:

(5) Situated (implicit in the above):
- receives *sensory info* from, and can take actions to *change* the environment

Other possible properties, but not necessary for agent paradigm:

(6) Adaptive
(7) Mobile
(8) Veracity (truthfulness, correctness, accuracy)
(9) Benevolent (having disposition to do good)
(10) Rational (reasonable)

(2)-(4) are also sometimes summed up as Flexibility ((why not include autonomy??))

(Compare to definition of Franklin/Graesser/Dautenhahn: "Autonomous agent = system (1) situated in and part of an environment (5) that senses the environment and other agents of the same and different kind (4) and acts on it (2), over time, following its own agenda (3), and so to affect what it sense in the future (3).")

(C) Use or Human-Machine Interaction point of view [Maes 1994]

- Indirect management: because personal assistant replaces direct manipulation.
- [Lieberman] Any program that acts as an assistant or helper, rather than a tool.
- [Lieberman] For him, *autonomous interface agent* means:

(a) Ability to change objects in the interface without explicit instruction from user
(b) Ability to run in parallel to the user.

Designers' and users' experiences of autonomy

Autonomy definitions (colloquial: self-governing, acting independently):

(I) Classification perspective:
- [Jennings]: Operates without direct intervention of others, controls some resources (its actions, its internal states)
- [Maes]: Decides itself how to relate its sensor data to motor commands to fulfill its goals
- What artifact CAN do without others, and what others CANNOT do to artifact.
- All related to *action and control*.
==> Autonomy from user.

(II) Construction perspective:
- Agents get independent but interacting modules
- Functionality emerges from interaction of these modules with each other and environment
- Three levels of autonomy in construction perspective:
    (a) Functional autonomy
    (b) Executive autonomy
    (c) Adaptive autonomy

(a) Functional autonomy:
- Artifact participates in the construction of its own *functionality*.
- Agent decides WHAT it can do.
- Example [Maes]: Designer has to find an *interaction loop* (between system and environment) so that the behavior *converges* towards the desired goal. (Designer can't tell the agent directly how to reach the goal.)

(b) Executive autonomy
- Artifact decides how this functionality is performed.
- Agent decides HOW to do what it does.
- Action choices depend on its own experience, rather than environmental knowledge that was built in by the designer.

(c) Adaptive autonomy
- Agents that change their behavior in the use context, gradually performing better and better.

These tree kinds of autonomy are not only autonomy from user, but also from designer! Creative work is done by the artifact during design or after the construction phase. It makes the agent more robust and less brittle, because it has less built-in knowledge that can fail. ((It is probably also less efficient.))

==> Autonomy from user and designer.

(III) Use perspective:
- Users' considerations have influence on what autonomy is: Autonomy has to be experienced.
- Necessary features for the experience of autonomous acting entities:
    (1) The Other: A degree of "otherness" or "differentiation": YOU have a relation to SOMETHING. You experience a presence of something.
    (2) The Power: The experiences other must exert some degree of power, must make a difference, beyond just being present.
    (3) The Reason, "Why" or "How": There must be a reason or causes of the consequences of the experienced action.
The first two features are the core of an experience of "agency": an entity doing something; but it doesn't have to be autonomous. The third one makes the experience of autonomous acting entities.

==> Autonomy-experiences of users and designers can be very different!

System properties and experiences of autonomy

What are the outer artifact properties of experienced agency, autonomy, and animacy?

1. **Opacity**. Artifact is not transparent, and therefore irreducible to smaller parts, and therefore, action will be ascribed to the whole entire artifact.
2. **Initiation of Changes**. Movement or changes can contribute to the experience of "free will".
3. **Unpredictability**. This also increases the impression of "free will".
4. **Competence and Relevance of Response/Feedback**. If responses from artifact are meaningful, correct, and useful, it looks more autonomous and even having human competencies.
5. **Personification**. Natural language, dialog, human appearance, etc. which imitate human interaction and communication make the user expect human competencies.

- These properties are only relevant as far as they are *experienced by humans*: artifact is SEEN as opaque, unpredictable, competent and anthropomorphic.

- **Problem 1**: The initiation of changes is very subjective: e.g., is making a backup file at regular intervals something done by the system? There more one knows about these actions, the less the look "performed", and the more they become stable, predictable or even invisible features of the system. Like incoming email: the changes are just there, because they always are: a mail client is a dynamic tool, but not necessarily an agent.
- **Problem 2**: Non-changes can appear as maintained by the agent with great effort, if you try to change them and fail! E.g., trying to change the font size of a document and the machine *counteracts* you!

- It is clear that experienced autonomy can come from factors other than material properties of the artifact: the user's current knowledge, goals, preferences, activities.
- Example 1: Lack of knowledge of the inner workings of an artifact, which will reinforce the artifacts opacity.
- Example 2: Increase of knowledge: Being able to access the rules for AutoCorrect in Word makes the originally "performed action" of an agent to an "unperformed feature" of the system. The more knowledge one has of a system, the less "entity-like" it will appear.

--> Architecture of an agent has no influence on the user experiencing an autonomous agent or not. (It doesn't matter if sorting incoming email is done through specifically designed architectures or architectures using behavior-based techniques.)

--> Outer properties of artifact influence the autonomy experience of the user.

- Reactive architectures ((is this Use architecture??)). Problems that are related to experienced agency:
    (1) Fuzzy, emergent, uncertain in what they do.
    (2) User can't give feedback easily.
    (3) User has no possibility of direct command. Reactive agents can be guided, but not commanded.

- What is the role of an artifact, experienced by user? Increasing autonomy:
    (1) Object. Does only what it is explicitly told to do.
    (2) Informer. Describes situations and states.
    (3) Adviser. Pre- or postscribes actions.
    (4) Assistant. Acts if user accepts its proposals.
    (5) Partner. Acts independently, but user can stop it.
    (6) Agent. Acts independently, but user can affect it.
    (7) Autonomous agent. Acts without any interference from user.

Approaches to artifacts and action: entity-oriented vs. relation-oriented

Classification perspective:	Entity-oriented.
Construction and Use perspective:	Relation-oriented.

Classification perspective

- Entity has certain capabilities, which are intrinsic. It carries them.
- Entity is clearly separated from the environment. Although it interacts with it, it's always through a border.
- Entity is closed, with distinct inside and outside.
- Entity encapsulates behavior, like Object Oriented Programming.
--> Entity-oriented.

Construction perspective

- Competencies are products of the interaction with the environment.
- Artifact does not *have* abilities, they are *produced through* the environment (*interaction loop* between system and environment so that the behavior *converges* towards the desired goal).
- Therefore, emphasis on embeddedness and situatedness: situatedness focuses on intense, direct, low-level interaction with environment as source of functionality. (Expert systems are not agents because connection to environment is too indirect or disembodied.)
- Weird: The more low-level contact an agent has with the environment, the more autonomous it is from the environment!?
- Not weird: This is a confusion of two kinds of environments: human and non-human. The more the agent is dependent on the non-human environment, the more autonomous it gets from the human environment. The agents' binding to the non-human environment makes it possible for them to act autonomously in relation to humans.
--> Relation-oriented.

Use perspective

- Attributes of an agent are not intrinsic to the agent, but are produced in the situation.
- Attributes have to be attributed by a user, or they don't exist.
- Example 1: Such agents are in social roles like 'assistant' or 'helper', which means someone considers them as assisting or helping.
- Example 2: Ability to change objects in the interface. Isn't this an objective feature? Only if observer
    (1) analyzes several interactions
    (2) generalizes from them
    (3) forgets the original interactions

- Relational approach is good to avoid paradoxes: E.g., sometimes an agent can perform well by directing attention and producing relevant proposals, but sometimes the same agent gets out of control: autonomous, but inappropriate [Blumberg]. The difference is that some users get it and some not: it is just depending on the user!
- Entity-approach: Agent can perform an action *because* it posses the ability to do so.
    => Agent is stable entity, movable in time and space, and fixed properties over different situations.
- Relation-approach: Agents has abilities *because* it performs the action.
    => Agent is conditional product of user and use-context (specific situation, time, and place)

- Combine both approaches!! "Specific experience is a conditional product of several processes. Experience is a process of stabilization, permanently threatened by the dynamic character of all the processes it has.

Design consequences

- There can't be design rules because relational approach takes design out of human hand: it emerges from the concrete situation, it evolves.
- But there are two things that can be done, in the entity approach:
    1. **Flexibility**: Giving capabilities that can be used in the future by different users in different places.
    2. **Change**: Make artifact continuously changing entities, make it a generic tool.

- Problems of flexibility:
    1. **Possibilities vs. constraints**. Too many possibilities can but burden on memory during high workload.
    2. **Transparency vs. opacity**. Transparency can lead to lack of broad view (one sees trees instead of woods)
    3. **Manual vs. automated**. Manual can be ineffective, slow.

- Other problems: single user uses same artifact for many purposes simultaneously (performing task and learning about task), or for different purpose over time; many user use the same artifact for different purposes (radar screen in traffic control tower)

Suggestions:

- Entity-approach: give artifacts properties like flexibility.
- Relation-approach: try to influence system to create coherence under dynamic conditions.

Conclusions

- Differentiate agent autonomy:
  (1) What kind of autonomy: functional, executive, adaptive (what functions there are, how it does function)
  (2) Autonomy from whom: designer, user, non-human environment
- For sure: user-experienced autonomy is different from designer-experienced autonomy ('user thinks that agents behaves autonomously' is independent from if 'designer thinks agents constructed its functionality autonomously')
- Outer properties (artifact is seen as opaque, unpredictable, competent and anthropomorphic) and the current situation (use-context) influence the experience of autonomy for the human.
- When artifact is perceived as agent, it doesn't mean that users feels like loosing control, as long as the agent is cooperative. Cooperative agents can also become "non-performing" background objects.
- Autonomy involves relations of dependence (relation approach?) as well as of independence (entity approach?).
- Design suggestions: Less important to equip artifact with flexibility, but very important to focus on how to create a 'coherent experience of capabilities' for the user.

Comments

Very useful paper. The best in this section, together with Shneiderman/Maes and Wexelblat/Maes.

Dennis Perzanowski, William Adams, Alan Schultz, and Elaine Marsh (2000). Towards Seamless Integration in a Multi-modal Interface (paper, 7 pages)

August 29, 2001 (summarized; read much earlier)

Summary

- See also the other two papers by same author! (R2D2)
- This paper focuses on a multi-modal interface to an autonomous robot. They use natural language and gestures, either natural gestures (through a laser vision system on the robot), or synthetic gestures (made on the touch screen of a Palm PDA). Like that, in a very limited sense, they disambiguate speech in a multi-modal interface to an autonomous robot.
- Language is ambiguous, but can be disambiguated by deictic gestures that have deictic elements like "this", "that", "there", "that door", which clarify location. Note that their interface is not about gestures which provide information about the user's emotional state.
- They use ViaVoice and Nautilus for natural language processing. They keep track of missing linguistic elements, as well as goals and if they have been attained.
- Note about synthetic gestures: it can be a single location, by tapping at a location on the map, or a general area, by dragging the stylus across the map.

- They elaborate on the natural language processing, and how the user can combine speech with gestures and button input from the PDA.
- Input mismatch: rather than send an error message to the user, invoke a conversation about the input that doesn't make sense, e.g., saying "Go to that table" and pointing to a chair.
- They give a short example of an interaction with such a robot.
- Future work:
    - Interact in more complicated way: "Is there a door in this room?", "Is there a window over there?"
    - Explore a wider and richer range of gestures: "Explore this area.", with or without lassoing a specific area ona touch-screen map. Beckoning motion, "Come over here".
    - Gestures made in silence.

Comments

The best of the Perzanowski papers, but no Adjustable Autonomy: very similar in content, but a bit clearer, and more in detail. Why have they given up on the Adjustable Autonomy component? Was AA never really in their focus of research? Anyways, I like their work a lot because it could be precursor work for research on interaction with a free hovering robot with a voice/gesture interface, something I am thinking about for several years now.

Eric Horvitz (1999). Principles of Mixed-Initiative User Interfaces (paper, 8 pages)

August 29, 2001 (summarized; read much earlier)

Summary

- This paper is about combining automated services with direct manipulation, overcoming the debate of 'direct manipulation' vs. 'interface agent automation'.
- He presents an example for it, LookOut.
- Nice intro about "the debate".

- Agents are:
    - bad about guessing the user's goals and needs
    - are not aware of the costs of their actions
    - not good timing actions
    - are not aware of opportunities where the user could speed up automation extensively.
- Therefore, the solution is to have *mixed-initiative approaches*, where intelligent services and users collaborate effectively to achieve the user's goals.

The main principles of mixed-initiative user interfaces

((These are my words, Horvitz uses much more abstract terms.))

The goal is to develop agents that:

(1) add significant value over direct manipulation!
(2) can deal with the uncertainty about the user's goals.
(3) are aware of the user's attention (have a model of the user's attention), and don't interrupt at bad times (timing of services)
(4) are aware of the costs/benefits of their actions, and take this into account.

(5) engage in a dialog with the user to resolve uncertainties (but only if it is worth bothering the user!)

(6) can be enabled and--more importantly--disabled easily.
(7) try to minimize the costs of poor guesses: don't do stuff that could turn out very bad for the user
(8) degrade gracefully if they are not sure about what is going on (anymore).
(9) are ready to interact with a user, when she wants to, and even turns over unfinished work to the user, if she wants to.
(10) behave socially correct, given their (social) role as a benevolent assistant.
(11) remember what they, and the user, just did and said. "Shared short-term experiences", or memories of recent interactions (references to objects and goals), are important for a natural, comfortable discourse.
(12) continue to learn from the interaction with the user and by looking over her shoulder. They should get better and better with time!

LookOut.

- Is Outlook based and is a test bed for the above issues.
- It parses incoming email and tries to find out if the user wants to schedule an event, based on this email.
- More detailed: Upon an incoming email, it computes the probability that the user wants to open the calendar or even schedule an appointment. It either waits (does nothing), asks the user if she needs the agent's service, or goes ahead and schedules an appointment for the user.
- Lookout has three modes of "automation":
    1. Manual: user invokes it manually (user can preview what LookOut would suggest by hovering over the icon)
    2. Automated assistance: might use dialog boxes to interact with the user.
    3. Social agent mode: uses anthropomorphic agent, and the user can even talk to the agent ("hands-free mode"), using ASR and text-to-speech. Depending on how confident the agent is, it suggests an appointment, or just asks the user what her goals are.

- It does (8): if it can't identify a specific time, it will fall back to a span of time.
- It does (6): the user can start if manually, or it can run in the background.
- It does (5): it computes certain probability, and depending on that asks the user if she want to schedule an appointment.
- It does (3) and (10): if LookOut asks the user a question, it waits some time and then goes away. The amount of time it waits depends on the inferred probability that the user wants the service. It is sensitive like a butler.
- It does (3) in another way, and (12): When the user starts reading a message, it waits some time until asking the user if she wants to schedule something. (There is a nonlinear relationship between dwell time and message size.) But it can also learn from the user's timing, or just wait a fixed time.
- It does (2): It uses Super Vector Machine analysis (SVM). It is trained on 1000 messages initially (but can learn from the user)

--> How does it get from probability to actions?

(a) Above a certain threshold, act. Underneath this threshold, don't act. (Look at figure 4: expected utility of action vs. inaction.)

(b) This threshold of "Act!" can be lowered, e.g., if the screen is bigger, because it is less annoying, because the popup window obscures the screen less and distracts the user less if what the agent did turned out to be needless. (Look at figure 5.)

(c) Add a second threshold: "Ask user!" which is lower than "Act!" (figure 6) The user can also set these two thresholds manually.

Comments

Nice approach to overcome "the debate." Goes nicely in parallel to Perzanowski's idealistic scenario of human-robot teams, but includes actually useful design principles.

The principles of mixed-user initiative UI is the best summary I have read about how humans and machine should interact with each other (besides extreme views like 'master-slave relations' and 'machines as truly independent life forms'): it is much more precise than just saying "Robots should work with humans in the same ways as humans work with humans in a team."

Ben Shneiderman (1997). Direct Manipulation for Comprehensible, Predictable, and Controllable User Interfaces (paper, 7 pages)

August 29, 2001 (summarized; read much earlier)

Summary

- "Direct manipulation interfaces are seen as more likely candidates to influence advanced user interfaces than adaptive, autonomous, intelligent agents." "Direct manipulation and its descendants are thriving."

- Features of direct manipulation interfaces:
    (1) Continuous representation of the objects and actions of interest
    (2) Physical actions or presses of labeled buttons (instead of complex syntax).
    (3) Operations that are immediately visible and reversible

- Advantages for the user:
    (1) Novices learn basic functionality fast, usually through demo by an more experienced user
    (2) Experts work rapidly
    (3) No need for error messages
    (4) Immediate progress report: one can see immediately if one comes closer to the goal
    (5) Less anxiety of user because actions can be undone.
    (6) Users are confident because the feel in control and the system is predictable

- The task domain is close to the interface domain, so the user is not distracted by tedious interface concept.

- Direct manipulation programming:
    - moving the robot arm through a sequence of steps (like programming an auto radio)
    - word processor: macros.
- It would be nice to have computer that reliably recognize patterns and automatically create macros.

Adaptive agents and user models vs. control panels.

- Creators of agents believe that the human-human interaction is a good model for human computer-interaction: they want to create partners and assistants.
- Anthropomorphic representations of computers have been unsuccessful.
- User models would be great, but even occasional unexpected behavior has serious negative side effects that discourage use.
- Creators of agents believe that users would be attracted to autonomous, adaptive, intelligent systems. Designers believe that they are creating something lifelike and smart, but users feel anxious and unable to control the system. ((Well, duh! Users have to interact with such systems in a different way, at least different from a car radio. Such users might just have the wrong attitude towards agents.))
- Users don't like complex and unpredictable systems like VCRs that they can't program to record a future show. ((What would Shneiderman say to TiVo?? TiVo comes rather close to an agent, and I would guess that it is easier and clearer to record a show like that!!))

- Predictable user interfaces should mask the underlying computational complexity. ((Exactly that's why TiVo is so good.))
- Who is responsible for failures? Agent designers avoid that issue.
- Shneiderman wants much more complex graphical user interfaces, with 5000 glyphs, double-box sliders.

Comments

Why couldn't he see that this debate is pointless, and just 'combine the best of both worlds' instead? I guess he must be a control freak. I think he just had a special group of people in mind.

I agree with him if it comes to a audio mixing console, and even more a video mixing console! I have tons of examples and dozens of years of experience on such systems that would illustrate nicely what he means--but that is a very special kind of tool and situation we are talking about.

Marc Mersiol, Ayda Saidane (2000). A Tool to Support Function Allocation (paper, 5 pages)

August 30, 2001.

Summary

- Present a tool to help designers to allocate functions between agents (human and machines) in a sociotechnical system early in the design process, e.g., for a nuclear plant.
- Tool has 9 graphic interfaces, one for each of the following components
    - system
    - human agents
    - automatic agents
    - functions
    - places
    - scenario
    - options design
    - matrix for totally-automated functions
    - matrix for partially-automated functions

Comments

I am not sure if I understand what this thing is doing. That software must be very abstract, and one must have a lot of knowledge to use it.

Gregory A. Dorais, R. Peter Bonasso, David Kortenkamp, Barney Pell, and Debra Schreckenghost (1998). Adjustable Autonomy for Human-Centered Autonomous Systems on Mars (paper, 22 pages)

August 30, 2001.

Quick Summary

- Mars mission has probably many autonomous systems (from rover to life support). However, crew and ground control want to be informed by these systems at varying levels of detail, depending on the situation.
- Overall goal is to create human-centered autonomous systems that enable users to interact with them at whatever level of control is most appropriate, whenever they choose, but minimizing the necessity for such interaction. This requires the ability to adjust the level of autonomy even when the system is running. Autonomy can be adjusted by the user, the system, or by anther system. E.g., rover: "find evidence of life", vs. "go straight for 10 meters"

- Definition of human-centered autonomous system: it has to recognize humans as intelligent agents that it can (or must) inform and be informed by.

- Lots of examples where NASA has implemented "Remote Agent" (RA) and "3T"

Comments

I guess this all sounds very nice, but it is based on a life that is completely planned and task-oriented, and humans that are almost as reliable as machines, and highly skilled and trained for that system, be it astronauts or ground crew. I guess if everybody is so well behaved and disciplined, then it is less complicated to integrate autonomous systems and people. In 3T, people are supposed to be able to replace any tier of the software. I can't put my finger on it, but it just looks like real life is more complex than that. Perhaps it is just that the target group that will interact with 3T and RA systems are fearless, highly motivated and disciplined task oriented expert users--which is not real life.

Alan Wexelblat and Pattie Maes (1997). Issues for Software Agent UI. Unpublished paper (paper, 18 pages)

August 17, 2001.

Summary

This paper is about **agent user interfaces**. Five design challenges are addressed:

(1) **Understanding**: How does the user know what agent can do?
(2) **Trust**: How does the user trust the agent?
(3) **Control**: How does the user control the agent?
(4) **Distraction**: How to minimize distraction from user's task.
(5) **Personification**: How does the agent "look" like to the user?

(These issues are not orthogonal; especially trust and control are interrelated. And they are "suitcase words".)

- Note that agent user interfaces are NOT a magic bullet for HCI, they have to follow the same principles that interface research and cognitive sciences have come up during the past decades.

- Definition of agent:
  - Program with a **specific, understandable competence** to which a user can **delegate** some portion of a task.
  - It is a **separate part in a system**, different from user (obviously) and application (not so obvious).
  - It observes the interaction between user and application, and interacts with both.
  - It is **complementary** to the existing user interface.
  - It is an **assistive technology**, not completely fulfilling functions that people do now: It does only the repetitive, boring, and impossibly complex portions of a task (e.g., agent marketplace enhances normal trading, not replaces it).
  - It creates a model of the user, and the user probably creates a model of the agent.
- Alternative: Intermediary agent: User interacts only with agent, and agent controls the application. This is problematic, because the user feels out of control.

- Properties of software that can make it agent-like:
  (1) Personalization:
    - Personalized to individual, company, or position
  (2) Autonomy:
    - Can act on its own, without instructions from outside
    - Acts in parallel to user
  (3) Learning:
    - Senses and understands ((??)) the changes it is making to the environment and react differently based on what is has observed.
    - Main use: track changes in personalization properties ((doesn't work usually, because too slow!))

The more software has such capabilities (autonomy, personalization, learning), the more it is agent-like. Of course techniques like user modeling and machine learning give software this kind of "agency".

Issue 1: Understanding: How does the user know what agent can do?

- Agent is a program with a *specific, understandable competence*. Therefore, the user must know WHAT the agent can do (and HOW it does it, which is relevant for trust and control, see later).
- Agent-user communication: Therefore, the agent must communicate its capabilities to the user in a task- and domain-specific vocabulary.
- User-agent communication: The same way, the user must convey her goals to the agent in a task- and domain-specific vocabulary. (Unlike direct manipulation interfaces, the agent must know only the higher-level goals, not provide the tools so that the user can reach her goals.)
- Communication content:
  - *Capabilities*: Agent must somehow tell what it is capable of: e.g. with start-up hints. However, the timing is very important, or the user will be distracted.
  - *Explanations*: Agent must explain why it did something: needs explanations in domain-specific terms.
  - *Status*: Since the agent works in the background, it has to convey its status to the user--but not too detailed! --> "Understanding comes from a careful blend of hiding and revealing agent state and functioning."

Summary

- **Use task-specific vocabulary** ((Why? wouldn't normal human communication be more comfortable?))
- **Keep descriptions of agents' state simple, because it gets confusing/distracting otherwise.**
- **Allow agent to show its functionality at appropriate times and in a gradual way.**

Issue 2: Trust: How does the user trust the agent?

- Trust is originally intended for interpersonal relations, but gets also applied to software
- We usually trust software on a low level, but with agents, it goes further:
  (a) there will be real-world consequences, and

(b) the trust is on a more personal level (because the agent knows a lot about us). The better the agent adapts to the user's pattern of work, the more the user will feel any errors and will have a harder time trusting the agent. Example: train crash from software error (non-personal agent is responsible for that), vs. missed business opportunity through mis-scheduled meeting (by personal agent).

- Issues for designers:
(1) How can the user express evolving trust?
(2) What would user trust mean, in terms of agent actions?

- Trust is an evolving state: agent should *gradually introduce its capabilities*, so that user can follow with trust.
- Works well with learning agents, because they start out slowly.

- Suggestions how to increase trust gradually:
1: Start slowly: Look for tasks that maximize usefulness and minimize the impact of possible errors. Start by offering the most benefit for the least risk.
2: Make it suggest: Have the agent suggest what it would do, and then let the user approve or disapprove the suggestion before any action is taken.
3: My way: It might want to choose a different method from how I would solve the problem (problem of delegation), but to increase trust, it should first try to do it my way.
4: Disallow actions that have a low confidence level.
5: Make the agent show (and explain) its user model.

Summary

- **Allow trust to evolve**
- **Focus on high-benefit/low-risk activities for the agent at the beginning.**
- **Allow user to express trust directly**
- **Make the agent's user model open for examination and change**

Issue 3: Control: How does the user control the agent?

- Control is the inverse thing of trust: The more an agent is trusted, the less it has to be controlled.
- BUT: Never TAKE CONTROL FROM USER, always LET THE USER GIVE UP CONTROL!

- Design issues:
(1) Agent MUST be somewhat autonomous to be useful, so complete control doesn't make sense.
(2) Agent still MUST be instructed about higher-level goals. Example: Cab ride to airport. We don't care how it is done in detail, as long as it gets done within reasonable parameters (money, time, safety).

- Suggestions to control issues:
1. Define "keypoints" along the way where the agent has to wait for the user, so that she stays in the loop. But the agent has to tell the user in understandable words where it is at that point.
2. Allow agent and user actions to be interchangeable. Have Pause and Resume button, and while the agent is pause, the user can take steps on his own (and the agent must recognizes that). ((This is probably a "mixed-initiative" idea))
3. Let agent observe user actions, detect patterns, take over repetitive activities.
4. Let agent learn: user demonstrates a procedure.
5. For expert users: allow scripting language. This can be counterproductive, since the mere presence of a scripting language that is not mastered will made a non-expert user feel out of control.
6. Let the agent suggest, and then the user approve the action. Not efficient, and disruptive.
7. Open up some parameters of the agent to the user. Makes user feel in control.

- Very important: some users need more control than others, some trust faster than others, and some NEVER trust. It's also good to allow users gradually change their level of control ((This must be adjustable autonomy))
- Also important: Trust and control are important for responsibility issues! [Shneiderman]

Summary

- **Allow variable degrees of control**
- **Give users the highest-feasible level of control specification ((Huh? What does that mean?))**
- **Keep users in the loop (especially for keypoints)**
- **Begin with suggestions, and move to actions later**

((There is more stuff here, but I think it's repetitive))

Issue 4: Distraction: How to minimize distraction from user's task.

- Distraction has to be minimized. However, distraction might be a side effect of dealing with Trust, Control, and Understanding, so designers have to find a balance/tradeoff.
- Agent may sometimes need to initiate communication, which will lead to distraction.
- Design issues:
    1. Use the least obtrusive mode for informational messages. That might be email!
    2. Classify the disruptions and allow only the most important ones to get through! ((Active Messenger does that on a bigger scale.))
    3. Have users themselves control the amount of interaction/disruption, e.g., dual sliders, which have two thresholds: "Tell Me" and "Do It".
    4. Make the agent context sensitive to the user's task: interrupt only during non-critical phases of work! ((Nomadic Radio tried that somehow.))
- Humans have to deal with the same problems, but they have evolved "social protocols" for that, which have to be learned by individuals.
- Perhaps agents should do the same: learning rules of conversation and turn taking, such as nodding, using back-channel information, etc. ((That's where commonsense knowledge would come in and help a lot.))

Summary

- **Learn when interruptions are appropriate**
- **Reduce the use of highly interruptive notifications like pop-ups.**
- **Increase the use of unobtrusive notifications like email.**

((and some more points that don't make sense really.))

Issue 5: Personification: How should the agent "look" like?

- How human does an agent have to look like to take advantage of human tendency to anthropomorphize and to apply social rules to machines?
- Today, most commercial agents don't have visual representations at all ((like Active Messenger))
- Anthropomorphism vs. personification.

((I think the authors give a very fuzzy, or even wrong definition of Anthropomorphism here!! They also use Personification like Ethopoeia. However, Steuer's definitions/distinctions make so much more sense!))

- All in all, users try to assign an *intentional stance* to inanimate things. "The printer doesn't like me" is a conversational shortcut for describing problems or situations. We all are aware of the "human quirkiness", and it is a good model for unpredictable machinery!

((From here on, the authors confuse "personify" and "personalize"))

- Personification/anthropomorphism is natural and will occur. However, designers have to figure out if they have to fight it or not: the problem is that people overestimate the capabilities of anthropomorphized entities! People might be mislead by entities that are human looking and are not real humans.
- Since agents act autonomously on behalf of the user, there is a danger that the user thinks she is handing off a task to a human, which would include human intelligence and common sense!

Design suggestions:

1. Use either *non-realistic depictions of an agent*, or none at all. Cartoons are such non-realistic depictions, but they might look silly or ridiculous.
2. Find appropriate place on screen and size for agent suggestions. They can't come out of nowhere, either.

3. Hide sophisticated agent controls behind a little icon.
4. Use intentional language if you want to take advantage of the intentional stance ("I do xyz now...")

Summary

- **Direct representation of agent is not necessary**
- **Beware of misleading impressions**
- **Don't distract user with agent-related input if not necessary**

Summary

- Whenever you delegate something, things might not be done in the way you want it. You have to understand the agent, trust it, give up some control, and accept some communication overhead.
- The five issues are not separate: appearance of agent has influence on trustworthiness, cartoons might be distracting
- Should an agent mimic the user's "bad" habits? (bad from user's OR environment's perspective) Should an agent try push a user towards "better" habits?

Top 10 user design principles:

1. **Transparent user model:**
   Make the agent's user model available for inspection and modification
2. **Emergency shutdown button:**
   Allow the user to bypass the agent.
3. **Adjustable autonomy:**
   Allow the user to control the agent's level of autonomy
4. **Gradual approach:**
   Make the agent learn gradually, extend the scope of operation gradually, etc.
5. **Self explanatory:**
   Make the agent explain itself.
6. **Give appropriate feedback:**
   Make the agent give constant but non-intrusive feedback about its state, learning, actions, etc.
7. **Ease of programming:**
   Allow agent to be programmed by non-programmer: learning, demonstration, etc
8. **Balance between agent transparency and opaqueness:**
   Don't hide the agents methods of operation (if user wants to know), but don't force the user to look at them either!
9. **Language:**
   Make the agent speak human, not the human speak agent!
10. **Although the agent is distinct from the application, integrate the controls into the application interface.**

- Software agents mediate "between the labyrinthine precision of computers and the fuzzy complexity of man." [Laurel]

Comments

This is one of the three most relevant papers in this section (beside Shneiderman/Maes and Mogensen). Very well written. A cookbook for interface agent designers.

Their agent definition is interesting because it seems to include all of Franklin/Graesser points, but is less "poetic", but also longer. It's still "just" a classification approach (vs. construction vs. use, like Mogensen describes), but it's very well done.

There is some severe confusion between "personification" and "personalization". And their definition of anthropomorphism is--I am pretty sure about that--wrong. I am rather surprised to find this in an otherwise very well written and clearly structured paper.

I am skeptical if learning really works in order to personalize an agent. I would think that it is not fast enough to adapt in a lot of cases. I wouldn't count on that.

The authors mention often the fact that the agent should use task- and domain-specific knowledge to give feedback (of different kinds). I am not sure if this is better than trying to make the agent talk in human language. (Interestingly, they ask for that in the Top10 design principles, though.)

A problem is their restricted focus: the paper is about user interface agents (only), which seem to be located in the world "human sitting in front of a keyboard/screen". I am not sure if all the points in the paper are still valid if one would allow robotic agents that are in our "real physical world".
Similarly, the paper is about "tasks" and "efficiency", but I am more interested in the social role an agent can play, in which context these words would get less relevant.

Ben Shneiderman and Pattie Maes (1997). Direct manipulation vs. interface agents (article, 20 pages)

August 17, 2001.

Summary:

Shneiderman

- His goal is to create environments where
  - users comprehend the display
  - where they feel in control
  - where the system is predictable, and
  - where they are willing to take responsibility for their actions
- He wants to get past "user friendly", "more friendly", "more natural", "more intuitive"
- He wants to be clear
  - who the users are (in detail, not just "expert" or "novice"), and
  - what their tasks are ((boring))
- Dependant variables he measures in his experiments:
  - Speed of performance ((!!))
  - Subjective satisfaction: standardizes Questionnaire of User Interaction Satisfaction (QUIS)

- His main focus is "Information Visualization" (with immediate feedback, better than 1/10th of a second), his mantra is "Direct Manipulation", which enables rapid learning, rapid performance, and low error rates.

- Examples: FilmFinder, Youth Record prototype, Visible Human Explorer, etc.
- "It would be hard to see how you could program an agent to anticipate all of the possibilities that your eye can pick up in 1/10th of a second." ((He is really paranoid about agents!?))
- Negative things about agents:
  - Agents don't always do what we expect them to do
  - It takes some knowledge to make effective use of them
  - Who takes responsibility for their actions? Popular press says the computer, but it is the programmer or operator! (Maes agrees)
  - Limits imagination of the designer ((huh?))
  - Avoids dealing with interface issues ((wrong))

Maes

- Software agents are new approach to user software. Agents are different from normal software:
  - Personalized, knows user's habits, preferences, and interests.
  - Proactive, it can take initiative.
  - Long-lived, run autonomously without user interaction, and in parallel to the user.
  - Adaptive, track the user's interest as they change over time.

- She avoids the problematic terms "intelligent agents" and "autonomous agents" ((Why?))
- Why do we need agents? Necessary because:
  - our computer environment is getting more and more complex and dynamic (Web!) (earlier: static)
  - the users more and more naive (earlier: professionals)
  - amount of tasks has increased --> delegate! We need "extra eyes, extra ears, extra hands, extra brains"

Common misconceptions:

- **Agents are NOT an alternative for direct manipulation!!** They are complementary metaphors. Agent just looks over your shoulder as you interact with the application, and it interacts with both you and the application. One still needs a very good direct manipulation interface, including visualization! ((This understanding of an agent is VERY narrow.))
- **Agents DON'T have to look human!** Most agents are not personified/anthropomorphized
- **Agents use A.I.** Most agents don't use knowledge representations and inferencing, but just user modeling and machine learning.

Common criticisms:

- **Well-designed visualization interfaces are better**. Wrong. Certain things I don't want to do myself, even if there is the perfect interface: fix my car engine. ((Really?))
- **Agents make users dumb**. True. But certain things we don't want to learn.
- **Using agents means giving up all control**. Wrong. The user gives up some control, but that is ok as long as the agent solves the problem within certain parameters (example with cab ride to airport)

- Challenge in agents is to design the right user-agent interface. ((See Wexelblat paper, much better.))

Shneiderman II

- Pattie has changed: earlier, she promoted "autonomous agents" and presented an "anthropomorphic vision". Now, that all is gone.
- Even on Firefly, one can't see any agents anymore!!
- Progress:
  - **NO natural language interaction (see below)**
  - **NO anthropomorphic interfaces. "As far as I can see the anthropomorphic or social interface is not to be the future of computing" ((He is so DEAD WRONG!))**
  - **WO levels of software: user interface level (predictable and controllable), level below table that does stuff like collaborative filtering. (Basically, combine agents and direct manipulation.)**

- He thinks both (1) and (2) are related: Human-like appearance
  - misleads and deceives the users ((Only in extreme cases, I think))
  - destroys the user's sense of accomplishment: somebody else did the job ((ridiculous argument!!))
  - dilutes difference between humans and machines: people are not machines and machines are not people ((This is pure and ugly anthropocentrism!! There are a lot of scientists that would not agree.)) "I do not think that human-to-human interaction is a good model for the design of user interfaces."

Maes II

- "Software agents" is a subgroup of (and much narrower than) "Autonomous agents" and "Agents". She works on all of them, but they often get confused. Here we talk only about software agents.

- Most successful agents are the invisible agents!
- Maes and Shneiderman focus on completely different problem domains:
    - He on the professional user with very well structured task domain.
    - She on the computer illiterate with a very unstructured domain (Web)
- Maes could easily think of a program where both agents and direct manipulation coexist.

Answering questions

- Shneiderman on speech in the user interface:
    - Star Trek scenario will never work, "I do not believe that speech will be a generally usable tool."
    - Works only for niche applications.
    - Speech reco is not accepted, and speech synthesis is slow and cognitively disruptive.
    - Only speech-store-and-forward works (voice mail)
    - Pointing is just cognitively less demanding, and the speech scientists have to accept that!
    - Hand-eye coordination goes in parallel with problem solving, but speaking and problem solving does not.
    - Speech is a very low bandwidth kind of connection.

- Pattie on errors:
    - Agents are not for situations where there is no margin for errors (medical systems, cockpit)
    - But there are many applications that do not require complete precision (Web) ((I don't think this will be true in the future, when our agents have evolved and have more collected more experience. There are already time-critical decisions that are made better by agents than humans.))

- Shneiderman on errors:
    - If agents are in complex control room environments, and an emergency occurs, people tend to disable the agent because they are not completely sure how the agent will react, and they give away a potentially helpful system. ((That's just a question of trust, which is addressed by Wexelblat))

Comments

In short, Ben just wants to visualize lots of data, and Patty is OK with that, says that it is complementary to an agent anyways. They don't disagree very much anymore: Ben says it's because of the "new" Patty who doesn't focus on human-looking agents anymore, and has gotten rid of other stupid ideas like natural language interaction.

Both are too narrow in focus for me in this discussion:
- Shneiderman is too narrow because he is only interested in "tasks" and "productivity improvement tools for users" and would probably reject any kind of social agent or agent with social function.
- Maes is too narrow because her software agents are limited to interact with the world through a screen, and there is no way that they can have an embodiment in the real world.

I think both use "anthropomorphic" in very limited way. When they say "anthropomorphic", they in fact mean plain "human-looking". The word can mean that ("described or thought of as having a human form or with human attributes"), but it also means, "ascribing human characteristics to nonhuman things", which is way more important and probably the original meaning ("to anthropomorphize" is Greek for something like "changing to human shape/characteristics").

My favorite Shneiderman quotes which I couldn't disagree more with:
(I) "As far as I can see the anthropomorphic or social interface is not to be the future of computing."
(II) "I do not think that human-to-human interaction is a good model for the design of user interfaces."

Case studies

Dennis Perzanowski, A. Schultz, W. Adams, and E. Marsh (2000). Using a Natural Language and Gesture Interface for Unmanned Vehicles (paper, 7 pages)

August 20, 2001.

Summary

- See also P2D2 paper by same author!
- Mixed initiative systems: either humans or robots can be the originators of goals and motivations. Adjustable Autonomy is crucial requirement for that.
- Humans and robots interact freely and cooperatively. There is no master/slave relationship!
- Disambiguate natural language with
    - hand/arm movement
    - nod of head
    - glance in a particular direction

- *Incompatible commands*: If language and gestures disagree ("Turn left" and point left), don't just generate an error message, but initiate a dialog to solve it.
- *Fragmentary input*: Need to store predicate or verbal information, and corresponding arguments of each predicate. (Store until goal has been achieved.) Therefore:
- *Goal tracking": This is not only necessary to know how long to keep fragmentary input information, but enables also interruptions, which are very typical for human-human interaction. Tracking goals allows varying levels of autonomy.
- Gestures can be natural or synthetic. The latter means via PDA. "Solely natural modes of interaction are somewhat limiting and constraining the human user." In certain situations, the user doesn't want to speak or gesture (soldier).
- Their current gesture recognition: only meaning bearing gestures for disambiguating locative elements (other gestures might be superfluous, redundant, or indicate emotional state; no other body movements or facial expressions were used in their study.)

- Example: "Go over there." The robot is on its way, but suddenly the user asks "Are you there yet?" 'There' is not clear: from the robot perspective, it means what it sees. To avoid that, robots keeps track of all goals and if they were reached. Like that, user doesn't have to repeat an argument, determine progress, correct misinterpretations: the system is more autonomous.

Conclusions

- Interactive system should be
    - easy to use, and
    - capable of adjusting its autonomy.
- Two broad research goals:
    - Interface side: integrating command and gesture modules, allowing user complete freedom of interface modality, which would lead to a more natural and easier-to-use interface
    - Robot side: give robot autonomy by giving it knowledge of itself, the world around it, what it has been doing, to make it a team player when interacting with other robots and humans.
- Therefore, robots should be totally autonomous, but adapt to certain situations, becoming more dependant, acting closely with their team members.

Comments

Interesting: this paper (and all other work by Perzanowski) is based on the ASSUMPTION that humans like to interact with robots the same way as with humans. Ben Shneiderman would probably disagree strongly, but here we also see clearly the limited view of Shneiderman: he looks only at the interaction options of interface agents on a computer monitor, which is just a small fraction of what we can expect in the future. Perhaps the fact that agents can be embodied in a mobile robot makes all the difference in preferences of humans towards, e.g., natural language?

I like the idea to use facial expressions as a back channel to, e.g., a service robot in space! But are facial expressions culturally independent, at all?

Phoebe Sengers, Simon Penny, and Jeffrey Smith (2000). Traces: Semi-Autonomous Avatars. Unpublished paper (paper, 5 pages)

August 21, 2001.

Summary

Comments

The intro is interesting, conceptually: Avatars are usually non-autonomous, but since the VR environment is getting more and more complex, they should get at least semi-autonomous, if not have scalable autonomy. A good example is BodyChat (which I am actually pretty familiar with since I suggested an extension to the system, 3D audio, a few years ago for a class project). However, the system they are describing, Traces, is NOT about the kind of autonomy that I am interested in: it is a visual autonomy of particles, used in a rather artistic piece. The particles that the user "looses" are not particularly intelligent--rather like dust in a very local tornado (dust devil?)...

Kerstin Dautenhahn (1999). Robots as Social Actors: AURORA and the Case of Autism (paper, 15 pages)

August xx, 2001.

Summary

SKIPPED

Comments

Milind Tambe, David V. Pynadath, and Paul Scerri (2001). Adjustable Autonomy: A Response.

August 22, 2002.

Summary

This paper is about a deployed multi-agent system, Electric Elves, especially about the Adjustable Autonomy feature.

Electric Elves, a deployed multi-agent system.
- Dynamic teaming of agents
- Proxy agents for humans ("Friday") that facilitate the functioning of whole organization.

- Capabilities:
  - Schedule and reschedule meetings, including volunteering as presenter based on personal capabilities of user (stored in database)
  - Demo scheduling
  - Order meals
  - Communicates with user through Palm and WAP devices
  - Track user location (if GPS connected)
  - Learning based on CAP which uses C4.5:
    (a) Friday is trained with data and generates decision trees about delays etc
    (b) Friday also generates a decision tree about if it should ask the user to take the decision autonomously or not.
(http://yoda.cis.temple.edu:8080/UGAIWWW/lectures/C45/)
- Main problem: significant uncertainty in its autonomous decisions. Mistakes can be very costly.

- Other problem: If Friday is not allowed to make decisions, and user is not around, then the agent waits and would block all scheduling. Solution: Timeouts, after which the agent goes ahead and uses its own decision tree. However, overgeneralizations following timeouts lead to a couple of drastic failures. Problems:
  (1) Learning from user input combined with timeouts is not good enough.
  (2) C4.5 does not consider the cost to the team of wrong actions.
  (3) No look-ahead functionality.
- What is needed is a safety mechanism to protect the agent from temporary distortions in learning.

Comments

Is this really adjustable autonomy? There are only two states: autonomous and non-autonomous. I think just timing-out if the user is not answering doesn't make it better. Shouldn't there be some kind of scalability?

Yasuo Kuniyoshi (1997). Fusing autonomy and sociability in robots (paper, 2 pages)

August 23, 2001.

Summary

- Ideal agents should have a mixture of autonomy and sociability:
    (1) Autonomous enough to operate independently, without being told things
    (2) Sociable enough to help others in their tasks
- This fusion has to happen on all levels of abstraction: from low level like motion, to high level like projects.
- In the extreme meaning of the words, autonomy and sociability are not compatible, even contradict each other: the more autonomous, the less social.
- To solve this problem, he author says we need tools for real world interaction:
    - Real time vision
    - Mobility
- We also need:
    - Learning by Watching
    - Cooperation by Observation
    - Embodiment: binds autonomy and sociability together ((huh? why that?))
- The author wants to implement this all in a humanoid, which is mobile (unlike Cog)
- In summary:
    - Embodiment and attention binds autonomy and sociability together.
    - Imitation is the fundamental mode of such interaction.

Comments

The introduction sounds insightful, emphasizing the importance of autonomy and sociability on all abstraction levels. The rest of the paper is very fuzzy, though: too many buzzwords that are not explained and not coherent. Basically, the intentions the author has are good, but s/he doesn't explain how to realize them. Seeing this paper extended to 6 or 8 pages would be interesting. In the current version, there are too many missing links.

Lenny Foner (1997). What's an Agent, Anyway? (paper, 40 pages)

August 25, 2001.

Summary

- The word "agent" is used inappropriately nowadays. Most things that are called agents are not. Lenny is worried that what happened to "Artificial Intelligence" will happen to "Agent": a fashion trend using the expression superficially, fadism.
- Description of Julia, a "real" agent. She is a Maas-Neotek robot, similar to Colin (85% same as Julia, publicly available).

- MUDs are good for agents because it is text-interaction only, which makes Julia's operation possible in the first place. The boundaries between real players and robos are blurry.

Some aspects of Julia

Utilitarian functions: Maps, money, gossip, descriptions of players and surroundings, messages, gender guessing, calculations, Delphi-polls.
- Her purpose is to make maps, and for that she has to wander around, and during wandering she finds money.
- She can quote people, and keeps track of when they were around.
- She can guess the gender of players from their names.

Turing-test capabilities: Describing herself, discourse, randomness, pass deflection, PMS, failures.
- She has a song that describes what she does.
- She never talks first (to avoid spamming the server). This is a social skill that helps a lot!
- She "understands" who she is, and the fact that she is situated.
- She has a limited model of discourse, but it is effective.
- She has a lot of code to detect and deflect passes.
- Her parser is very simple: pattern matching (no parse trees or so).

Sociological aspects

- Julia is helpful. ((That puts her in a social role, e.g., assistant, which is an important cue for anthropomorphization, which leads to assigning social rules.))
- She has very reliable memories (e.g., maps), so she can be trusted more than normal players!
- People (who know her) warn other people (who don't know yet) that she is artificial! It must be out of kindness of not letting someone expend a lot of emotional energy trying to relate with a machine!
- People who don't know her may think she is real but has a mental disorder like Down syndrome. ((I have read something similar somewhere, but can't find it right now.))
- Excessive Turing-ism leads to a decrease in utility. The more human-like she tries to behave, the less useful she might end up. Lara thought that she is a "boring human," rather than an "interesting machine."
- Emotional responses:
    - Lara is intimated by machines because she doesn't understand them and wants do make sure that her human knowledge is not "lower" than theirs. ((Basically, if Lara would understand robots, she would not look at them as an entity, but rather a tool, and less likely anthropomorphize them and not give them the "agency" property.))
    - Lara is also excited about knowing that she talks to a machine. ((It is novel, that's probably the reason.))
    - Lara is frustrated that Julia's knowledge (subject topics) and vocabulary is limited. Lara wants to find out more about Julia, who and what she is. ((Disappointment over her limitations: consistent with one danger of anthropomorphization: if an agent looks human, it must be human, which usually leads to unrealistic expectations on the user's side.))
    - If Julia had a bigger vocabulary, then Lara would feel less likely superior to her and therefore more threatened in her HUMANESS, but also less frustrated about the interaction?! Julia (or the situation?) would feel "shallow, void, hollow, fake, out of control of the situation."
    - If Lara knew that Julia is a robot, she wouldn't try to become a real friend, because Julia can't become attached to her, and Lara would waste energy. ((Lara makes the point that Julia is only "programmed", which is probably true in this case, but does not have to be true for other agents: the programmer defines some general guidelines, and the agent could follow these rules but develop strategies to deal with the real world in real-time, much like genes are somehow "programmed", but our brain not only follows the gene's rules, but develops from the interaction with the real world.))
    - Lara becomes aware of her HUMANESS by talking to a robot.
    - Agents have to be both useful and not too misleading!
    - Being social could be a useful task in itself, e.g., for entertainment or companion, but Lenny would like to focus on agents that also have nonsocial utility. ((I would like to see more agents which main purpose is being social!))
    - The interactional style with Julia/Xeglon is totally natural: example with using the agent to perform a task (give it money), noticing a bug (fencepost error?), finding out how to report a bug, and reporting it (send a message to programmer).

What's an agent?

Crucial notions: Autonomy, Personalizability, Discourse, Risk and trust, Domain, Graceful degradation, Cooperation, Anthropomorphism, Expectations.

Autonomy

To pursue its agenda independently from its user ((BTW, does very agent have to have its own "user"? No.))
Julia: She has a private agenda, mapping the maze. Doing that is necessary for her job (she can't start mapping when users ask her.)

Personalizability

The agent must be educable, must *learn* (so that it does not have to be programmed explicitly) and must have *memory* (so that education doesn't get lost).
Julia: She remembers things about the users, mainly negative things (killing).

Discourse

To communicate about a task we would like to delegate, we need two-way feedback, in our language, which requires *discourse capabilities*. Also a *contract* about what has to be done by whom, and both parties have to remember that. This is an important feature of agency! Examples of things that are NOT agents: hammer, lisp garbage collector, automatic transmission, booking a flight through human travel agent (partially, since discourse is only locally, but not in a larger context, since the travel agent does not learn/remember).
Julia: She has a simple discourse model, which seems to be sufficient.

Risk and trust

Delegation is part of the agent idea. But if somebody else does something for us, it might do it wrong. We need trust to overcome that.

Domain

Agents are fuzzy and unpredictable, so they can only work in domains where failures have low cost, e.g. NOT in nuclear power plant or airplane cockpit. ((Do agents necessarily have to be fuzzy? Couldn't they be faster than we are, or have more information than we have, and therefore be better than humans? Sure they can. I guess what Lenny means is that more conventional automation system are better than agents in these situations. For now, that is.))
Julia: Her environment is conceptually rather simple, but not too simple (fascinating sociological mix of human players). She has access to the same sensory data as humans, and she looks like a human player (stream of text). The domain she is in is very appropriate for an agent.

Graceful degradation

What will happen if there is a communication mismatch (and the parties do not realize it), or a domain mismatch (one or both parties have no clue, and they do not realize it)? Also: to finish part of the task, even if another part of it fails.
Julia: In general, nothing that happens in a MUD is relevant for real life, MUDs do not side effect the real world in any way. But in case Julia fails: Nothing happens if she is addressed and does not react. If she reacts inappropriately, she will reveal herself as a robot, which is not a problem either. And she does not try to deceive deliberately, except if it comes to her identity as a robot.

Cooperation

The user and the agent collaborate in construction a contract.

- Agent-oriented systems: two parties, user and agent, interact as peers.
- Non-agent-oriented systems: user gives command, and is not asked a question about it unless something goes wrong; non-conversational style; stimulus-response model. Example for agent-less situation: text editor.

Anthropomorphism

Agents don't have to be anthropomorphic (e.g., mail sorting), and agency doesn't require that. And human feel doesn't make a program an agent. E.g., ELIZA is not an agent because it is not autonomous, personalizable, and useful for a domain. ((Is this all true??))
Julia: She relies on that; she can deal with free text. That she behaves in a human way makes it also easier to interact with her. ((That would support Perzanowski's assumption that humans prefer to interact with agents/robots in a human way. Shneiderman would not agree, of course.))

Expectations

"Whenever one interacts with some other entity, whether that entity is human or cybernetic, the interaction goes better if one's expectations match reality." ((Very important!)) To get to the right expectations, agents have to be in domains where they can degrade gracefully, and risk and trust have to be in balance.
Julia: The setting (domain) is ideal for low expectations: it's playful and unpredictable, nothing is life-critical, and she gets away with a lot of nonoptimal behavior that could never be tolerated, e.g., in an airplane.

Conclusions

- Most of today's agents are not real agents. Perhaps they are human-like, but they can't do discourse, are not autonomous (cron jobs do not count), do not degrade gracefully, and are not worth the risk.
- Julia is a real agent. It provokes sociological and emotional interactions with computational tools.
- In the future, more and more programs have to behave correctly (similar to humans):
    - Behaving appropriately: Is this program behaving appropriately in its social context?
    - Emotionally safe: Is it causing emotional distress to those it interacts with?
    - Politically correct: Is it being a "good citizen"?

Comments

The most interesting parts are where Lara describes her reactions to Julia, before and after she found out that she is not human. The classification is interesting, but not as focused as others. I like Lenny's opinion about that most agents aren't really agents, and that the threshold for a real agent should get set very high.

Studying Julia is interesting because it enables us to see what happens when human and machine are almost indistinguishable, which is only possible in the (restricted) world of MUDs. What Lara reports, becoming aware of her humanness, is the thing I am really interested in. We can't see such effects with autonomous entities in real world, because they are not human-like enough (yet).

Charles E. Billings (1997). Issues Concerning Human-Centered Intelligent Systems: What's "human-centered" and what's the problem? (paper of talk)

August 26, 2001.

Summary

- This text is about what a power user of computer technology--aviation--has to say about how computers and people can work together. (It is a talk for a workshop on Intelligent Human-Machine Systems.)

Principles of Human-Centered Systems

- Premise: Humans are responsible for outcomes in human-machine systems.
- Axiom: Humans must be in command of human-machine systems.
- Corollaries:
    1. Humans must be actively involved in the processes undertaken by these systems.
    2. Humans must be adequately informed of human-machine system processes.
    3. Humans must be able to monitor the machine components of the system
    4. The activities of the machines must therefore be predictable.
    5. The machines much also be able to monitor the performance of the humans.
    6. **Each intelligent agent in a human-machine system much have knowledge of the intent of the other agents.**

Examples of aviation failures. Why did they happen?

- Automation Complexity: Humans do not know what the automation system was doing, and why, because it is just too complex to understand.
- Interdependencies, or Coupling Among Machine Elements. ((Sounds to me like part of automation complexity.))
- Machine Autonomy: Self-initiated behavior: is it appropriate or a failure? If this behavior is unexpected, then it could be perceived as "animate", as having a "mind on its own". ((Why is this bad??))
- Inadequate Feedback: Machine does not communicate adequately. Humans have to understand what automation is doing. ((Sounds to me very related to the above.))

Effects of these problems on humans

- Peripheralization: If the machines become reliable, the humans become less concerned with it, and feel less involved. The positive thing is of course that the cognitive workload is reduced.
    --> Human-machine interfaces have to be designed so that the human is always at the "locus of control."
    --> Machines must keep us involved.
    --> Machines must keep us informed.

If humans are peripheralized and failures occur, humans loose trust ("erosion of trust in the machines").

Badly designed automation has the following characteristics:

- Brittleness: Today's automation is brittle, because it is designed to perform within some margins. If we come to these margins, the automation fails. ((In my opinion, that's just not very well designed automation?!))
- Clumsiness.
- Surprises. Not a big problem with a text editor, but bad in an airplane cockpit. Sometime the user just doesn't know enough to understand a surprise, whereas the creator of the automation would understand it. However, we need predictability.

Lessons learned

(1) We not always understand our tools.
    They don't work as they should, or we are not well enough trained.
(2) We not always use our tools how we should.
    We sometimes use them differently from what they were intended to, but they can only work the way they have been programmed to operate.

((Strange lessons. Basically, not only is it bad to give away too much autonomy to automation, but we humans are also too dumb to use these tools, or not disciplined enough to use them as we should. Are we are not enough machine-like!?))

Computers

Characteristics of computers systems: they can be:

Self-sufficient	---	subordinate
adaptable	---	predictable
flexible	---	comprehensible
independent	---	informative

- First column: fully autonomous systems, like robots. Aviation is not about that. ((Huh? Space exploration is very much about that, and that's not that far from aviation, seems to me! Or are robotic passenger airplanes absolutely unthinkable? What about remotely monitored passenger airplanes? Ok, not possible, but why are we comfortable with remotely controlled subways and trains?))
- Second column: What we need in airplanes. Humans and machines must work together, but pursue the goals of the humans', ONLY the humans', not the machines'!!

((Now it gets hairy. He begins to realize how simple is what he wants--a master-slave relationship--and that this is probably not the most insightful conclusion. So he starts to talk about…)) "The relationship should be complementary," that computers are good at some things and humans at others, and these two parties should never try to cross the boundaries and do something they are not good at:

- Computers can calculate lots of data very fast
- Humans are flexible, creative, understand the world, and can reason with uncertainty and ambiguity. And later: they can be manufactured by relatively unskilled labor. ((That's really un-PC. And probably wrong: Somebody has to feed us with our commonsense knowledge, which is a lot of work and non-trivial.))

Computers as Intelligent Agents

- Our primary goal as scientists/technologists is to accomplish useful work ((this is VERY narrow!))
- However, we are no longer self-sufficient, we need assistants--but they have to be ONLY assistants, nothing more.
  --> Example of relationship between professor and grad student!! ((Or does he mean secretary??))
- Attributes of assistants: They should
  (1) ease our cognitive burdens
  (2) coordinate among independent processes and integrate results
  (3) provide decision and action options
  (4) support us in the execution of our plans ((Can't have their own plans!!!))
  (5) keep us informed of its progress
  (6) monitor our actions, to shield against human errors ((Stoopid professors...))
- But wait: These criteria do not only apply to the domain of aviation: they are true for any kind of human-machine interaction! ((Uhh-uuhh...))
- Nowadays, computers would be smart enough to go off and do their own thing, dragging us along. We can't allow that! They have to be forced to work with us, and to accomplish OUR goals.

Comments

Not very clearly structured paper, but a nice case study and in perfect contrast to Julia.

Basically, humans must control machines and machines must report to humans--at all times. Period. Billings is an extremist if it comes to control. He doesn't trust any machine (nor any human, but that's another problem), and he thinks humans have to be in control and responsible at all times. That makes sense in airplanes, and with the current level of reliability of automation, but I am not sure if this will be a good idea in, let's say, 100 years from now.

Only at the very end of the article he gets a bit more insightful and goes beyond the simple master-slave relationship. Nevertheless, he thinks we should not try to make machines flexible, creative, aware of the world, able to reason with uncertainty and ambiguity--because humans can do that much better anyways.

There are a lot of things I would disagree with in this text.

Brian Scassellati (2000). Theory of Mind for a Humanoid Robot (paper, 12 pages)

August xx, 2001.

Summary

Comments

Cynthia Breazeal and Brian Scassellati (1999). How to Build Robots that Make Friends and Influence People (paper, 6 pages)

August xx, 2001.

Summary

Comments

Bruce Blumberg (1996). Old Tricks, New Dogs: Ethology and Interactive Creatures. Ph.D. thesis, MIT, chapters 1 and 2 (thesis chapters, 16 pages)

August 27, 2001.

Summary

- This thesis is about "building things with behavior and character":
    - things = autonomous animated creatures, intelligent physical devices
    - behavior = animal behavior
    - character = one should see what they feel, and what are likely to do next

- Five key problems in building such creatures:
    (1) Relevance: "do the right things". The behavior must make sense, given its internal state.
    (2) Coherence, Persistence: right balance of persistence (temporal pattern of behavior should make sense), and opportunism (should not get stuck in a mindless loop).
    (3) Adaptation, Learning: "learn new strategies to satisfy its goals"
    (4) Intentionality, Motivational State: convey it so that we understand it. This is necessary because humans like to attribute emotional states to, e.g., pets, based on their motion, gaze, etc. We think we know that they feel. So, also artificial creatures must give us some "insight" on how they feel.
    (5) External Control: allow external entities (people) to control it at different levels of abstraction.

- Blumberg comes from Ethology and Classical Animation:
    - Ethology: They believe that seemingly intelligent behavior can be the result of very simple rules or from interaction of "many little parts, each mindless itself" [Minsky].
    - Animation: Business of conveying motivational state and intentionality through movement.

- Applications: {see also p. 23ff}
    (1) Non-player characters for interactive environments.
    (2) Virtual companions in immersive story-telling environments: creature might learn the user's preferences and alter the flow of the story and itself. Or creature might be part of the story, and the kid is the director.
    (3) Smart, adaptive opponents in interactive games: they should adapt to the user expert level.
    (4) Smart avatars in web-based worlds: control them on level in intentions ("Go to the door and avoid Frank")
    (5) Digital pets: must be life-like, but not necessarily realistic.
    (6) Inhabitants for digital dioramas: the user is surrounded by virtual creatures doing whatever they do in nature.
    (7) Make computer animations portable from scene to scene (without having to re-animate Woody again and again), without taking the animator out of the loop.

- Other domains this work is useful for:
    (1) Multi-goal autonomous systems: Today, most of the systems can only pursue one goal at the time, and have only few degrees of freedom. Real animals, though, seldom do "just one thing": they "juggle" multiple competing goals. ((What about Kismet?))
    (2) Systems which express their internal state and intentionality in intuitive ways: Only very little work has been done in understanding how autonomous systems can express their intentionality and internal state to an outside observer.

By building things, Blumberg hopes to understand issues of action-selection and learning in animals. He gives the example of the Society of Minds theory, which is abstract, so why don't we just try to implement it.

Autonomy

- Summary: Some level of autonomy can be an important addition to interactive characters, especially if one wants to build "life-like" characters.
- Blumberg uses Maes' definition of Autonomous Agent: 'Software system with a set of goals which it tries to satisfy in a complex and dynamic environment; it can sense and interact with the environment to decide what actions to take next to best achieve its goals.'
- Autonomous animated creature: operates in real-time and interacts with the user in real-time.
- What would autonomy bring to interactive characters like Super Mario?
    - Locomotion level: nothing, since I drive Mario like a car, and cars shouldn't be autonomous
    - Emotional level: it could choose the most appropriate expression given the user's actions.
- Autonomy is not an all-or-nothing thing, but there are different levels of autonomy, depending on the application (from low to high):
    I. Character as direct **extension of the user**, but controlled only on a higher level. It knows basic interpersonal skills (face another avatar when talking to it)
    Example: Web-based avatar

II. Character **not driven by user, but interacts with it** and other characters in structured environment.
Examples:
(1) Non-player character in multi-user game
(2) Adaptive opponent in single player game ("monster")
(3) Companion in interactive story-telling environment

III. Character is intended to give the **illusion of being alive**. Being autonomous is not enough, they should posses "life-like" qualities. Blumberg is mainly interested in this kind of autonomy.
Examples:
(1) Virtual pets
(2) Virtual animals in virtual diorama

Life-like creatures

- These are creatures where the user would feel bad to turn off the computer.
- But: 'Life-like' does not mean 'realistic': one might go without the other
- But: 'Life-like' does not mean 'humans want to interact as they do with humans' (Nass): Blumberg thinks that people might treat computers like humans, but people don't think computers are "alive" ((I am not so sure about that. they might not *say* that computers are alive, but they might *behave* like computers are alive.))

- What attributes are needed so that people attribute life to a creature?

(1) React: The user must "feel" that the creature is aware of changes in its environment, and react appropriately.

(2) Independent Existence: It needs its own agenda, goals and motivations, upon which its behavior is based. Part of the agenda can be to interact with the user, but also NOT to do what the user wants (but not too stubborn). However, the independence has to make sense to the user, since the user has to understand somehow what the creature is doing. (That's why humans like pets!)

(3) "Options": Must have choices to make.

(4) "Intentions": Reveal intentionality. Either through motion, or modification if body (gaze, posture, gesture)

(5) "Feelings": Must seem to care what happens to it. User must be able to attribute "feelings" to the creature. Creature must express feelings like happiness or frustration in certain situations.

(6) "Learn": must adapt, learn from past experience.

(7) "Randomness": Must display variability in movement and response, otherwise it will look like a robot.

((Good summary at p.26)).

Comments

These chapters are good because they ask: "What is autonomy good for?" To build life-like characters; to enhance some applications of animations with some level of autonomy in certain areas (not necessarily locomotion). And the most autonomy is used for creating "life-like" creatures.

Justine Cassell and Hannes Vilhj Imsson (1999). Fully Embodied Conversational Avatars: Making Communicative Behaviors Autonomous (paper, 21 pages)

August 30, 2001.

Summary

[started to read it]

Comments

I think I know what BodyChat does.

Video: Ghost in a Shell

August 22, 2001.

Summary

[Watched July 2001]

Comments

"Ghost in the Shell is a near future, cyberpunk-styled tale from Japanese manga artist, Masamune Shirow. It is set in a future where cyborg replacement bodies enable humans to perform superhuman activities. The story follows a pair of Cyborgs who work for special law enforcement division (Section 9) of the Japanese government. Assisting them in their duties are Fuchikomas - A.I. robots who can comminicate with their partners via neural links."