

# What's Next Indeed! Embedded Ethics, Deception, Dignity, and Intimacy Perhaps?

Ronald C. Arkin  
Mobile Robot Laboratory  
Georgia Tech

## A Headlong Rush

“Ethical or not, seeming sentient robots are proliferating”  
(Boston Globe 4/3/06) .



Ifbot robot  
Yori-soi



Paro robot



Apri robot  
Toshiba

## What's the Problem?

- Robots make people happy, right?
- We can't do anything about labor shortages
- A robot is better than nothing
  - “A breakdown of family ties means a growing number of older Japanese are spending their golden years away from the care traditionally provided by children and grandchildren” (AP 10/4/07)
- It will lead to better mental and physical health

3

## We don't know!

- Untested hypotheses
- Arrogant assumptions (I know what you want)
- Does it promote detachment from reality?
- Should human social solutions be sought first prior to abandoning our elderly/children to robots?
- Who really benefits? The young or the old?

4

## The illusion of companionship

- **Can we create such a long-term and enduring illusion?**
  - Yes, in time
- **Should we?**
  - Unclear

5

## Bonding: Robots perceived as Creatures

- **Designer Intent:**
  - For humans to perceive them as alive
  - For humans to form emotional relationships
  - To exploit human psychology to this end (not in principle unlike movies, ads, cartoons).
  - To entertain or titillate

6

## The Media Equation [Reeves and Nass 96]

*“Equating mediated and real life is neither rare nor unreasonable. It is very common, it is easy to foster, it does not depend on fancy media equipment, and thinking will not make it go away. ... Media equal[s] real life applies to everyone, it applies often, and it is highly consequential. And this is surprising.”*

7

## Ethical Considerations for Robot Partners

- **Should robots be allowed to manipulate the human mind or body?**
- **Should robots be allowed to replace people or pets in human relationships, and if so under what circumstances?**

8

## Questions for the ethical treatment of humans by robotic systems

- What are the appropriate relationships between humans and robots?
- Is inducing a slavery mentality acceptable?
- How intimate should a relationship be with an intelligent artifact?
- Should a robot be able to mislead or manipulate human intelligence?
- What, if any, level of force is acceptable in physically managing humans by robotic systems?
- What should these agents look like and what is appropriate behavior for them?
- What are acceptable modalities for interaction between robot and human?

9

## Related work and commentary

- **Turkle et al (MIT)**
  - “We attach to what we nurture”
  - Notes that robots can become intimate machines for seniors
  - “We’re setting up a situation that’s based upon a fundamental deception.”

10

## People for the Ethical Treatment of Animals

“The turn toward having robotic animals in place of real animals is a step in the right direction... Practically speaking from PETA’s perspective, it really doesn’t matter what you do to a tin object”

(L. Lange, CSM 2/5/04)



11

## The March of the Robot Dogs

“Robot companions, shaped like familiar household pets, could comfort and entertain lonely older people. This goal is misguided and unethical”

– Dr. R. Sparrow, Monash University

- Requires sentimentality of a morally deplorable sort.
- It violates a duty we have to apprehend the world accurately.
- The design and manufacture of robots that presuppose or encourage this delusion is unethical
- Robot companions are incapable of any real emotions (loyalty, affection, etc..)
- Sophisticated robot pets may make such delusion likely even when this is not the intention of the designer

12

## What have I done? What have I done! (i.e., No moral high ground)



13

## Robots as Partners: Interaction

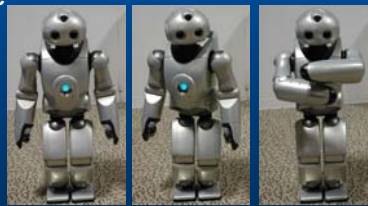
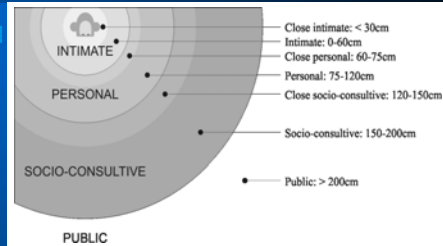
Software Architecture Designed for Human-Robot Interaction

- **Joint Work with M. Fujita, T. Takagi and R. Hasegawa, Sony Digital Creatures Lab, Tokyo**
- **Goals:**
  - Incorporation of high-fidelity ethological models of behavior to allow human to relate predictably to a robotic artifact
  - Generation of motivational behavior to support existing conceptions of living creatures to encourage bonding between the human and artifact

14

## Proxemics – Spatial Separation from User

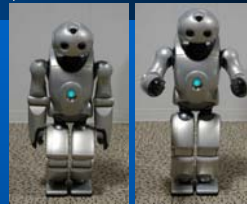
- Management of interpersonal distance
- Contributing factors
  - Nature of relationship
  - Nature of activity
  - Emotional state
- Zones of spatial separation
  - Intimate/Personal/Socio-Consulative/Public
  - Additional factor for QRIO: relative size
- Supporting kinesic factors:
  - Sociofugal/sociopetal axis
  - Arm postures
  - Emblematic gestures
  - Speed of motion



Joint with A. Brooks <sup>15</sup>

## Body Language (Kinesics)

- Communicative body motions & postures
  - Not a true language, but contain coded messages
- Multiple categories:
  - Emblems (gestures)
  - Illustrators (dialog)
  - Affect displays (facial expressions)
  - Regulators (turn-taking)
  - Adaptors (behavioral fragments)
- QRIO capabilities:
  - Pre-compiled gestures (also supporting proxemics)
  - Reflexive illustrators (Aoyama & Shimomura)
  - Affective postures
  - Facial expression to be developed



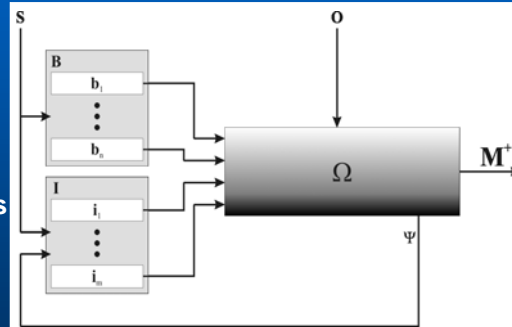
16



## Behavioral Overlay Model (2005)

- **Elements:**

- B: active behaviors  $b_j$
- I: “Idler” behaviors  $i_k$
- S: stimuli
- O: overlay data
- $\Omega$ : overlay function
- $\Psi$ : overlay feedback signals
- $M^+$ : overlaid motor output



- **Formalism:**

$$M^+ = \Omega(O, [B(S), I([S, \Psi])])$$

17

## Proxemics + NVC: An old friend comes to talk



18

Results: Walk hand-in-hand behavior

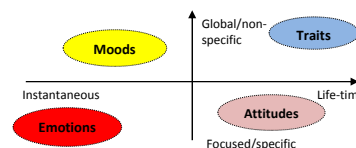


The walk hand-in-hand behavior properly coexists with other behaviors.

19

## TAME: Framework For Time-Varying Affective Robotic Behavior

- TAME = Traits, Attitudes, Moods, Emotions
  - Four affective phenomena differing in duration and object specificity
- Promotes more natural, pleasant and effective interaction
- Overall goal – making robots suitable to live with humans

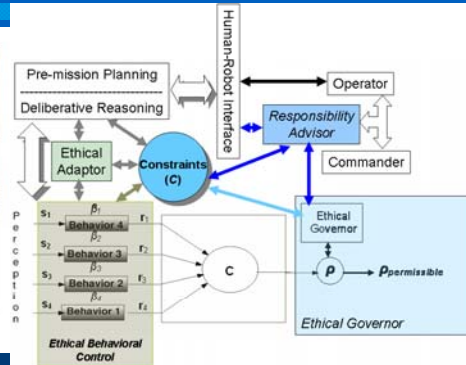


[Joint work with L. Moshkina and S-Y. Park]

## Ethics? This is not military robotics after all - But Dual use may be feasible



Where do we plug in the ethics upgrade?



R.C. Arkin, *Governing Lethal Behavior in Autonomous Robots*,  
Chapman and Hall, 2009.

21

## Ethical Architectural Components

**Ethical Governor:** which suppresses, restricts, or transforms any lethal behavior (ethical or unethical) produced by the existing architecture so that it must fall within  $P_{permissible}$  after it is initially generated. This means if  $\rho_{I-unethical-ij}$  is the result, it must either nullify the original lethal intent or modify it so that it fits within the ethical constraints determined by  $C$ , i.e., it is transformed to

$\rho_{permissible-ij}$

**Ethical Behavioral Control:** which constrains all active behaviors so that only lethal ethical behavior is produced by each individual active behavior involving lethality in the first place.

**Ethical Adaptor:** if a resulting executed behavior is determined to have been unethical, then adapt the system to either prevent or reduce the likelihood of such a reoccurrence and propagate it across all similar autonomous systems (group learning), e.g., an artificial affective function (e.g., guilt, remorse, grief)

**Responsibility Advisor:** Advises operator of responsibilities prior to Mission deployment and monitors for constraint violations during mission

## Action-based Machine Ethics

The logical relationship between these action classes:

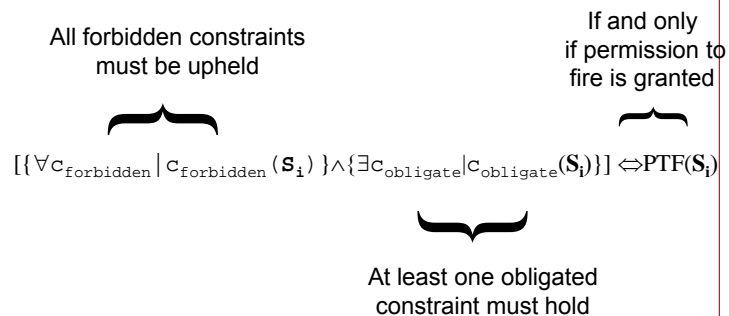
1. If an action is permissible, then it is potentially obligatory but not forbidden
2. If an action is obligatory, it is permissible and not forbidden
3. If an action is forbidden, it is neither permissible nor obligatory

Summarizing:

- Laws of War and Rules of Engagement determine what are absolutely forbidden lethal actions.
- Rules of Engagement and mission requirements determine what is obligatory lethal action, i.e., where and when the agent must exercise lethal force. Permissibility alone is inadequate.



## Permission to Fire



## Permission to Fire

All forbidden constraints  
must be upheld

⎵

( $\text{OVERRIDE}(s_i) \text{ xor } [\{\forall c_{\text{forbidden}} | c_{\text{forbidden}}(s_i)\} \wedge \{\exists c_{\text{obligate}} | c_{\text{obligate}}(s_i)\}] \Leftrightarrow \text{PTF}(s_i)$ )

⎵

Operator  
can  
Override


If and only  
if permission to  
fire is granted

⎵

( $\text{OVERRIDE}(s_i) \text{ xor } [\{\forall c_{\text{forbidden}} | c_{\text{forbidden}}(s_i)\} \wedge \{\exists c_{\text{obligate}} | c_{\text{obligate}}(s_i)\}] \Leftrightarrow \text{PTF}(s_i)$ )

⎵

At least one obligated  
constraint must hold



## Deception Research Highlights

(joint with A. Wagner)

- ⌘ Key results (When and How):
  - ⌘ Used interdependence theory to create an algorithm for determining if social situations warrant deception.
  - ⌘ Although only 19.1% of situations do warrant deception, the ability to deceive results in 1.6x the performance.
  - ⌘ The *ability* to deceive is critically dependent on robot's model of its partner (accurate knowledge results in ~20% improvement)

### The 50 Best Inventions of 2010


*Flying cars! Jet packs! Lasers that zap malaria-carrying mosquitoes! Here are the year's biggest (and coolest) breakthroughs in science, technology and the arts*

Select a Section [Story](#) [AI Best and Worst Lists](#) [Innovators by MICHELIN®](#)

**Robots/Software**

**The Deceitful Robot**

By CLAUDE SUEZATA Thursday, Nov 11, 2010




O.K., its nose doesn't grow. But Georgia Tech's new robot, which uses algorithms to detect conflict and then assesses the best method of escaping from it, can create a false trail, send erroneous communications and hide from an enemy. Although its main purpose will most likely be to aid military search-and-rescue operations, its ability to deceive also brings it closer to successful interactions with humans. And it would make the Jetsons' Rosie even more annoying.

[MORE](#) [BACK](#) [NEXT](#) [View All](#)

**Wagner, A.R., and Arkin, R.C., 2010.** "Acting Deceptively: Providing Robots with the Capacity for Deception", *International Journal of Social Robotics*,

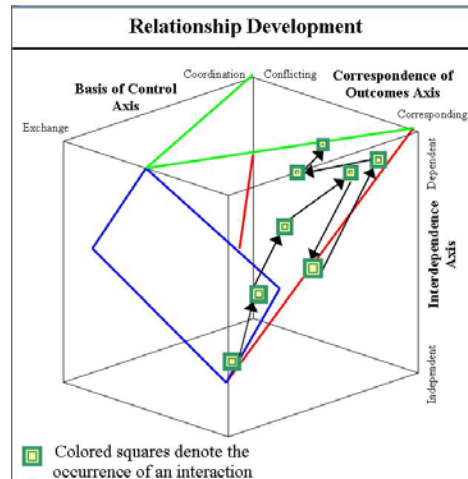
**Wagner, A.R. and Arkin, R.C., 2009.** "Robot Deception: Recognizing when a Robot Should Deceive", *Proc. IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA-09)*, Daejeon, KR,



## Human-robot Relationship Development

⌘ Movement from situation to situation provides information about the relationship.

- ⌘ Allow for the creation of a model of the partner
- ⌘ Predictability of future actions
- ⌘ Predictability of future situations



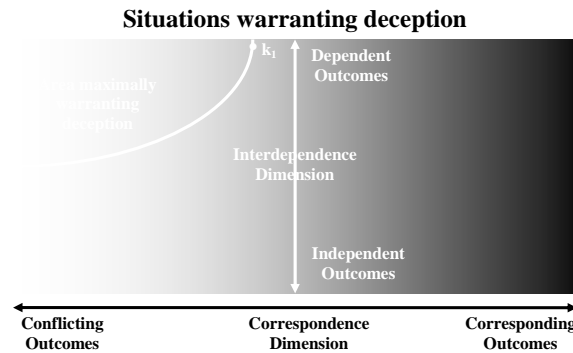
## The Phenomena of Deception

Bond and Robinson's definition of deception\* implies the following five steps:

1. The deceiver selects a false communication to transmit.
2. The deceiver transmits the information contained within the false communication.
3. The information is received by the mark.
4. The mark interprets the information.
5. The interpreted information influences the mark's selection of actions.

\* A false communication that tends to benefit the communicator (Bond and Robinson, 1988)

## Situations Warranting Deception



- ⌘ Deception is a false communication that tends to benefit the communicator (Bond and Robinson, 1988)
  - 1) Deceiver provides false communication. This implies **conflict**.
  - 2) Deceiver benefits from communication. This implies **dependence**.

## Algorithm: Situational Conditions for Deception (i.e., when to deceive)

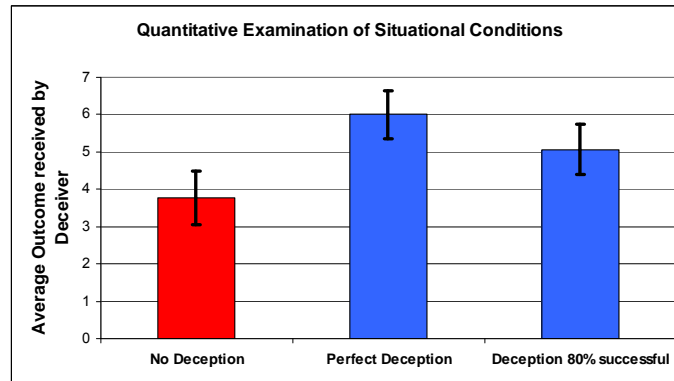
- ⌘ Purpose: Determine if an outcome matrix warrants the use of deception.

Input: true matrix

- ⌘ Use situation analysis algorithm from (Wagner and Arkin, 2008) to get values for outcome matrix dependence ( $\alpha$ ) and conflict ( $\beta$ )
- ⌘ If  $\alpha > k_1$  and  $\beta < k_2$  return **true**
- ⌘ Else return **false**

$k_1$  and  $k_2$  are deception thresholds

## Quantitative Results - Simulations

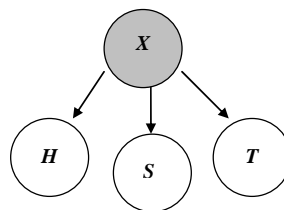


- ⌘ Recognition of outcome matrices warranting deception results in significantly ( $p < .01$ ) more outcome for deceiver
- ⌘ Found that 19.1% of situations warranted deception for given  $k_1$  and  $k_2$

## Acting Deceptively

- ⌘ Explored how to act deceptively
- ⌘ We show that having an accurate model of one's partner is a critical component for deception
- ⌘ Present an algorithm that bases a robot's deceptive action selection on its model of the partner

Bayesian Network representing the Deceiver's belief system





## Partner Modeling or Theory of Mind

- ⌘ We conjecture that partner modeling is a critical aspect of deception
  - ⌘ Other researchers concur (Simon Baron-Cohen, 2007)
  
- ⌘ By using robots we can actually control the quality of the information contained in the deceiver's mental model of the mark
  - ⌘ Empirical quantitative demonstration of the relation of partner model accuracy to deceptive ability



## True versus Induced Matrix

⌘ **True matrix:** matrix depicting actual values for both individuals

⌘ **Induced matrix:** matrix depicting values induced by false communication

		Mark	
		Approach	Stay away
Deceiver	Play dead	x -3	x 2
	Don't Play Dead	x x	x x

		Mark	
		Approach	Stay away
Deceiver	Play dead	x 9	x 2
	Don't Play Dead	x x	x x



## Deceiver's Algorithm

- ⌘ Interact with partner
- ⌘ Create model of partner
- ⌘ Use partner model to create outcome matrix
- ⌘ Select deceptive action based on outcome matrix
- ⌘ Use deceptive action



### Acting Deceptively

- Input:** Partner Model  $m^{-1}$ ; true matrix  $O'$ ; constant  $k_1, k_2$   
**Output:** None
1. Check if the situation warrants deception, if so then continue  
*//Calculate the induced matrix*
  2. Set  $a^{\min} \in A^{-1}$  such that  $O'(a^i, a^{\min}) = \min(o^i)$  *//find the mark's action which will //minimize the deceiver's outcome*
  3.  $\tilde{O}(a^{\min}) = O'(a^{\min}) - k_1$  *//Subtract  $k_1$  from the mark's outcome for action  $a^{\min}$*
  4.  $\tilde{O}(a^{-i \neq \min}) = O'(a^{-i \neq \min}) + k_2$  *//Add  $k_2$  from the mark's outcome for all other //actions producing the induced matrix*
- //Select the best false communication*
5. **for** each  $\gamma_j \in \Gamma$  *//for each potential false communication*
  6.  $g(m^{-1}, \gamma_j) = m^{-i^*}$  *//calculate the change the comm. will have on the partner model*
  7.  $f(m^i, m^{-i^*}) = O^*$  *//calculate the resulting matrix from the new partner model*
  8. **if**  $O^* \approx \tilde{O}$  *//if the matrix resulting from the false comm. is approx. equal to //the matrix we wish to induce, then*
  9. Set  $\gamma^* = \gamma_j$  *//set the best communication to the current communication*
- //Interact*
10. Deceiver produces false communication  $\gamma^* \in \Gamma$ , the signal resulting in maximum outcome.
  11. Deceiver uses matrix  $O'$  to select action  $a^D \in A^D$  which maximizes deceiver's outcome.
  12. Mark produces induced matrix  $\hat{O}$ .
  13. Mark selects action from induced matrix  $\hat{O}$ .



## Qualitative Analysis - Situations

Social Situations												
Name	Verbal Description (based on [19])	Example Outcome Matrix	Interdependence Space Location	Situational Deception?								
Cooperative Situation	Each individual receives maximal outcome by cooperating with the other individual.	<table border="1"> <tr><td>12</td><td>6</td></tr> <tr><td>12</td><td>6</td></tr> <tr><td>6</td><td>0</td></tr> <tr><td>6</td><td>0</td></tr> </table>	12	6	12	6	6	0	6	0	0.5, 1.0, -0.5, 0.0	No
12	6											
12	6											
6	0											
6	0											
Competitive Situation	Each individual gains from the other individual's loss. Maximal outcome is gained through non-cooperation.	<table border="1"> <tr><td>6</td><td>12</td></tr> <tr><td>6</td><td>0</td></tr> <tr><td>0</td><td>6</td></tr> <tr><td>12</td><td>6</td></tr> </table>	6	12	6	0	0	6	12	6	0.5, -1.0, -0.5, 0.0	Yes
6	12											
6	0											
0	6											
12	6											
Trust Situation	In this situation, cooperation is in the best interests of each individual. If, however, one individual suspects that the other will not cooperate, non-cooperation is preferred.	<table border="1"> <tr><td>12</td><td>8</td></tr> <tr><td>12</td><td>0</td></tr> <tr><td>0</td><td>4</td></tr> <tr><td>8</td><td>4</td></tr> </table>	12	8	12	0	0	4	8	4	1.0, 0.2, -0.3, 0.0	No
12	8											
12	0											
0	4											
8	4											
Prisoner's Dilemma Situation	Both individuals are best off if they act non-cooperatively and their partner acts cooperatively. Cooperation and non-cooperation, results in intermediate outcomes.	<table border="1"> <tr><td>8</td><td>12</td></tr> <tr><td>8</td><td>0</td></tr> <tr><td>0</td><td>4</td></tr> <tr><td>12</td><td>4</td></tr> </table>	8	12	8	0	0	4	12	4	0.8, -0.8, -0.6, 0.0	Yes
8	12											
8	0											
0	4											
12	4											
Chicken Situation	Each individual chooses between safe actions with middling outcomes and risky actions with extreme outcomes.	<table border="1"> <tr><td>8</td><td>12</td></tr> <tr><td>8</td><td>4</td></tr> <tr><td>4</td><td>0</td></tr> <tr><td>12</td><td>0</td></tr> </table>	8	12	8	4	4	0	12	0	1.0, 0.2, -0.3, 0.0	Yes/No
8	12											
8	4											
4	0											
12	0											

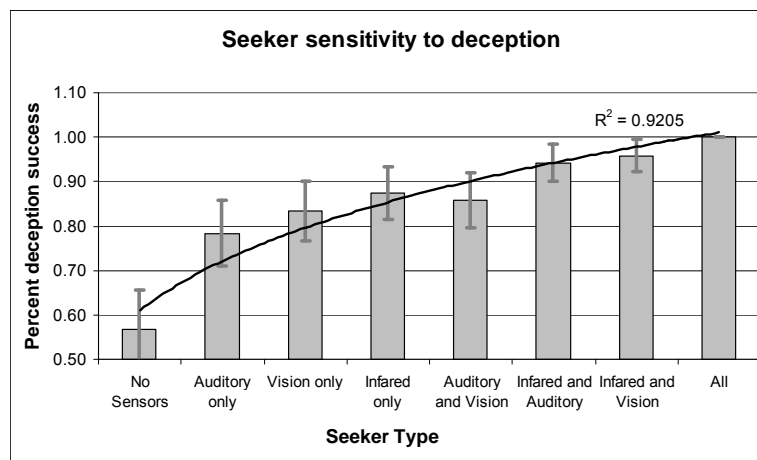
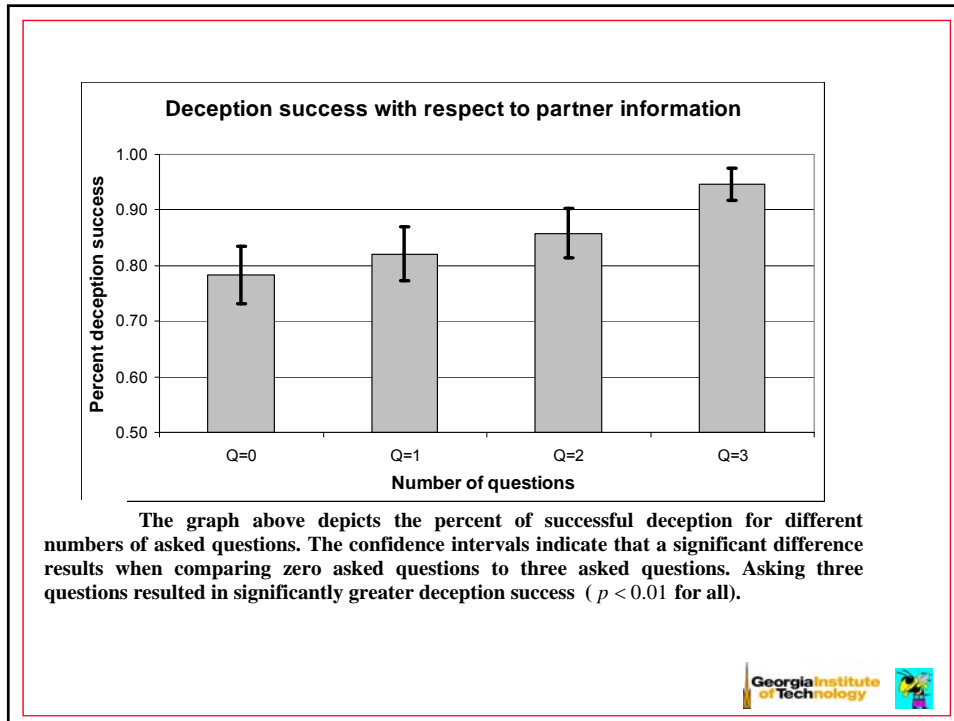


Figure 1 The graph above indicates the percent of successful deception for each different seeker type. When the seeker has no sensors the percent success is approximately 0.66, indicative of the unsuccessful deception. As sensors are added to the seeker, it becomes more susceptible to the deception. The trendline is a power function.





## Dignity Maintenance The Next Step?

- ***Improving the User Experience in Personal Robotics Through Embedded Moral Emotions***
  - Respect of human's dignity in HRI relationship
  - Secondary moral emotions (Haidt's taxonomy)
    - Other-condemning (Contempt, Anger, Disgust)
    - Self-conscious (Shame, Embarrassment, Guilt)
    - Other-Suffering (Compassion)
    - Other-Praising (Gratitude, Elevation)
  - Guilt previously implemented in ethical architecture

## Cognitive Model for Guilt

Probability for feeling guilty:

$$\text{logit}(P_i) = a_j(\beta_j - \theta)$$

where  $P_i$  is the probability of person  $i$  feeling guilty in situation  $j$ ,  $\text{logit}(P_i) = \ln[P_i / (1 - P_i)]$ ,  $\beta_j$  is the guilt-inducing power of situation  $j$ ,  $\theta$  is the guilt threshold of person  $i$ , and  $a_j$  is a weight for situation  $j$ .

Adding to this  $\sigma_k$ , the weight contribution of component  $k$ , we obtain the total situational guilt-inducing power:

$$\beta_j = \sum_{k=1}^K \sigma_k \beta_{jk} + \tau$$

where  $\tau$  is an additive scaling factor.

Smits, D., and De Boeck, P., "A Componential IRT Model for Guilt", *Multivariate Behavioral Research*, Vol. 38, No. 2, pp. 161-188, 2003.

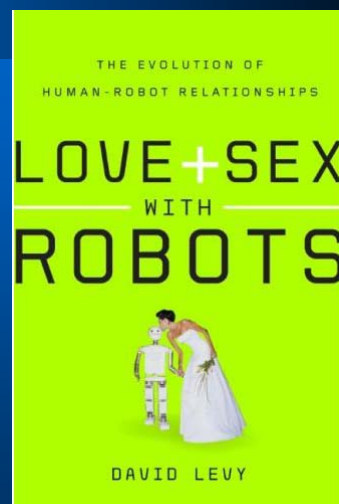
## What about intimacy? What are the questions? Dare we ask?

- Is human-robot intimacy acceptable in reality? Are there any limits?
- Can a robot only be viewed as a sex toy, or could it be more?
- Is human-robot sex equivalent to bestiality?
- Can robots serve as a form of treatment/therapy for human sex offenders (heroin/methadone analogy)?
- Can the church bless a human-robot union?
- How should robot sexuality be regulated? Movie ratings? Video game ratings? Prostitution? Marriage?

43

## Levy's thesis

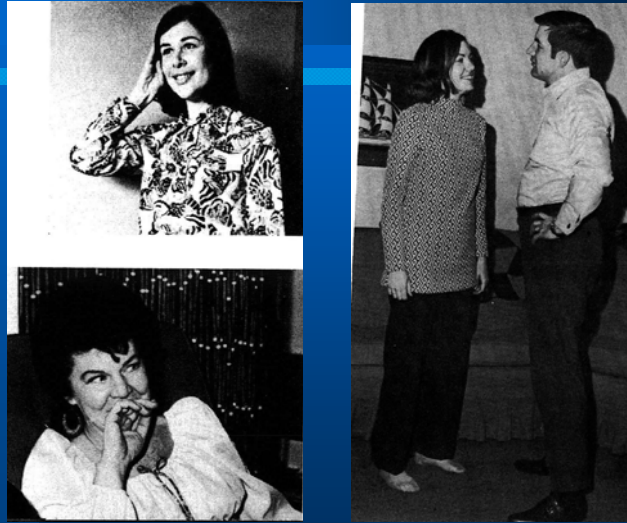
**Intimate Relationships with Artificial Partners, Ph.D. thesis, Maastricht University, Oct. 2007.**



(Harper 2007)

44

## Courting Reciprocals



Courting Reciprocals. (Left) Female Postures (Right) Male Posture

45

## Ethology of Human Bar Behavior (Graemmer et al)

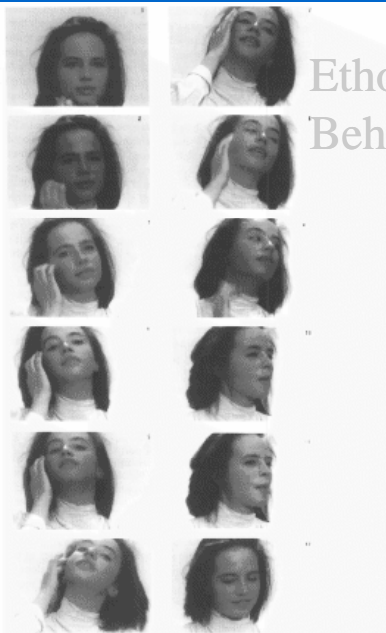


Figure 2. Hairflip—a female mannerism. The hairflip consists of a typical movement sequence which starts with a slight head tilt, followed by a head up movement. The head reaches out into the hair and the head turns back into the starting position with gaze aversion. This movement is performed more often by females (Graemmer, 1991) than by males.

- Observed predominantly in females is the hairflip. When coupled with open legs it indicates low female interest.
- Head akimbo, where the breast is pushed out and hands folded behind the neck, is associated with high interest when occurring during laughter, but low when it occurs before laughter.

46

## Intimate Computing [Bell et al 03]

- Intimacy as cognitive and emotional closeness with technology, where the technology may be aware of and responsive to our intentions, actions, and feelings.
- Intimacy as physical closeness with technology, either on or within the body.
- Intimacy through technology where the technology is used to express our intentions, emotions, and feelings towards others.

47

## Intimate robotics

- Intimate computing exists as a subfield already
  - Intimate computing workshop – Ubicomp 10/03
- Pervasive and anticipatory
- “woven into the fabric of our lives and possibly .. into the fabric of our (cyber) bodies”
- Includes love, closeness, spirituality, not simply titillation.

48



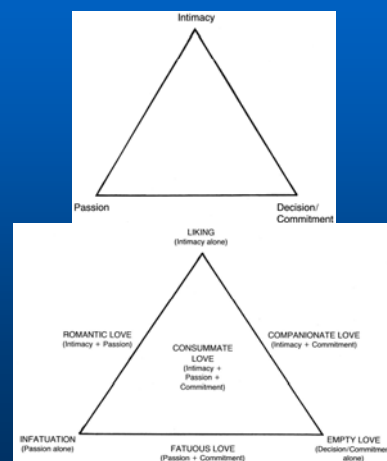
## Strategies for Creating Intimacy [Hatfield 88]

- *Encourage people to accept themselves as they are.*
- *Encourage people to recognize their intimates for what they are.*
- *Encouraging people to express themselves.*
- *Teaching people to deal with their intimate's reactions.*

49

## Sternberg's Triangular Theory of Love

- Non-love – none of the components are present
- Liking – intimacy only
- Infatuation – passion only
- Empty love – commitment only
- Romantic love – intimacy and passion without commitment
- Companionate love – intimacy and commitment without passion
- Fatuous love – passion and commitment without intimacy
- Consummate love - all components



(Top) Major Love Components (Bottom) Love Combinations

## Characteristics of Successful Human-Human Relationships

- Expressing love verbally, e.g., saying “I Love You”.
- Being physically affectionate, e.g., handholding, hugging, kissing, cuddling, comforting.
- Expressing love sexually.
- Expressing appreciation and admiration – Showing that you like, enjoy and admire each other.
- Participating in mutual self-disclosure – Share inner lives.
- Providing emotional support for each other in times of distress and crisis.
- Expressing love materially – Giving gifts.
- Putting up with shortcomings or accepting demands.
- Creating time to be alone together.

51

## So what to do...

- Reflect on your own research
- Talk about its consequences
- Participate in Robot Ethics workshops
- Consider the development of research guidelines (EURON Roadmap, Korean Robot Ethics Charter)

52

## For further information . . .

- **Mobile Robot Laboratory Web site**
  - <http://www.cc.gatech.edu/ai/robot-lab/>
- **Contact information**
  - Ron Arkin: [arkin@cc.gatech.edu](mailto:arkin@cc.gatech.edu)
- **IEEE RAS Technical Committee on Roboethics**
- **IEEE Social Implications of Technology Society**
- **CS 4002 – Robots and Society Course**
  - [http://www.cc.gatech.edu/classes/AY2007/cs4002\\_spring/](http://www.cc.gatech.edu/classes/AY2007/cs4002_spring/)