Carnegie Mellon

# Validation

Kathleen M. Carley
Carnegie Mellon University

Center for Computational Analysis of Social and Organizational Systems
http://www.casos.cs.cmu.edu/

Carnegie Mellon

## Definitions

- Validation – a set of techniques for determining whether or not a model is valid. Used for both internal validity, matching with other models, and matching with non-computational data.
- Calibration – a set of techniques for tuning a model to fit detailed non-computational data.
- Training – procedures for supplying data and feedback to computational models that can learn.
- Verification – a set of techniques for determining the validity of a computational model's predictions relative to a set of non-computational data.
- Docking – a set of techniques for determining the level of comparability or equivalence of two models.

Carnegie Mellon

# Methods and Levels for Validating a Computational Model

- Validation techniques vary in
  - Method
  - Level
  - Intensity
  - Purpose
- Similar approaches can be used for
  - Calibration
  - Training
  - Verification
  - Docking

Carnegie Mellon

# Validation Levels

- Internal validity - error free code
- Parameter validity - parameters match
- Process validity - processes fits
- Face validity - right type of things
- Pattern validity - pattern matches observed
- Value validity - values match
- Theoretical validity - theory fits

Carnegie Mellon

# A Caveat

- Computational modeling is sufficiently complex that a single individual in a single research period (e.G. 6 months to a year) can not build, analyze, and validate a computational model.
- Most models take multi-person years to build and analyze.
- Data collection and analysis from a virtual experiment often takes as long as a human experiment and requires statistical training comparable to that required for human experiments.

CASOS



Carnegie Mellon

# Face Validity

- Is the model a reasonable simplification of reality?
- Techniques to increase face validity:
  - Set parameters based on real data
  - Model a specific organizational or inter-organizational process
  - Show that others have made similar assumptions
  - Discuss model limits and how left out factors may or may not affect results
  - Don't over-claim model applicability

CASOS

Carnegie Mellon

# Illustration: Model & Reality

| Simulated Annealing | Organizational Strategic Adaptation |
|---|---|
| system | organization's CEO or central unit |
| state | organizational design |
| current state | current organizational design |
| temperature | risk aversion |
| accepting a cost increasing move | taking a risk |
| high temperature means accepting many cost increasing moves | liability of newness |
| move set | re-design strategies |
| heuristic optimization process | satisficing & BR process |
| minimize cost | maximize performance |
| cooling schedule | approach to becoming risk averse |
| proposed state | proposed new design |
| evaluation of proposed state | limited lookahead, anticipation of future |
| state evaluation | observed performance |

CASOS

Carnegie Mellon

# (Social) Turing Test

The model does the task it seeks to explain.
1) substitutability.
2) Turing test.
3) social Turing test.

Construct a collection of social agents according to the hypotheses and put them in a social situation, as defined by the hypotheses. Then recognizably social behavior should result.

Aspects not specified by the hypotheses, of which there will be many, can be determined at will.

The behavior of the system can vary widely with such specification, but it should remain recognizably social.

CASOS

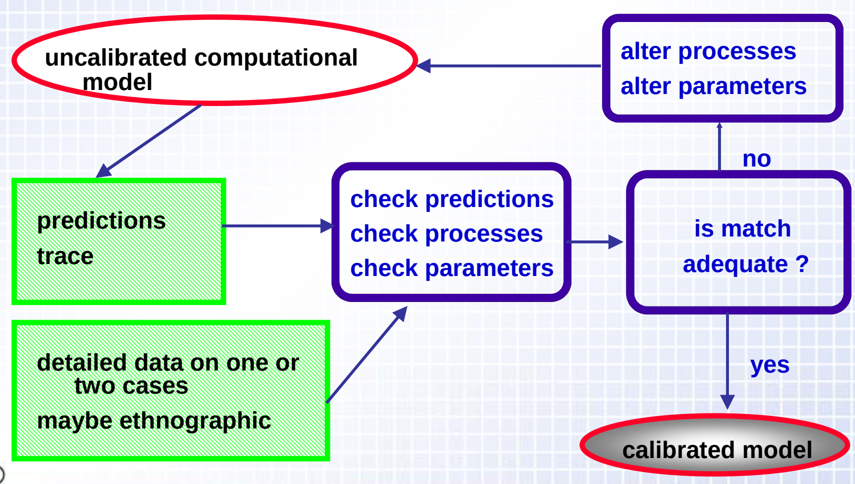
Iterative build-check process

# Calibration

- Calibration involves fitting a computational model to a set of data
  - May require programming (adding modules or new processes)
  - May require parameter setting
- Have available detailed data on one or more cases
- Calibration is often the only validation step carried out for emulations
- Calibration demonstrates that model can match non-computational data



# Calibration Cont.



Carnegie Mellon

# Locating Cases

- Ideally:
  - Use a set of cases that span the key categories you are concerned with
- Next best:
  - Choose 2-4 cases that represent typical behavior and 1 or 2 that represent atypical behavior
- In practice:
  - Most intellective models are not calibrated
  - Lucky to have even one case with sufficient detail
  - Often detailed case is a matter of opportunity
- Sources:
  - Archival data, ethnographies, participant observation

CASOS

June 13, 2006  11

Carnegie Mellon

# Verification

- Verification involves testing a computational model's predictions given a set of non-computational data
- Have available the results of a virtual experiment
- Have available non-computational results
  - May be archival, survey, experimental
- Verification is sometimes done on uncalibrated models
- Verification demonstrates that model's predictions match non-computational data

CASOS

June 13, 2006  12

Carnegie Mellon

## Types of Verification

- Level of verification:
  - Pattern - same trends are observed
  - Value - same values are observed
- For multi-agent models:
  - Group or organizational level - matches overall behavior of collection of agents
  - Agent level - matches specific entities behavior
- For stochastic models:
  - Point - on average behavior is the same
  - Distribution - distribution of results is the same
  - Detail match - one entire run is the same

CASOS

June 13, 2006  13

Carnegie Mellon

## Data for Verification

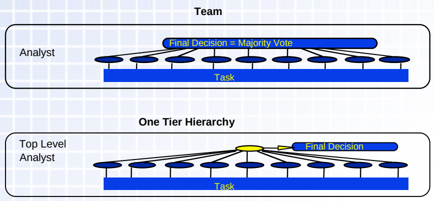
- Type: anything
  - May be archival, survey, experimental
- Quantity: high
  - Sufficient for statistical analysis
- Level of detail: low
  - Do not need the same level of process data that is needed for calibration

CASOS

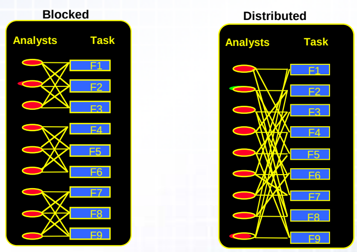
June 13, 2006  14

Carnegie Mellon

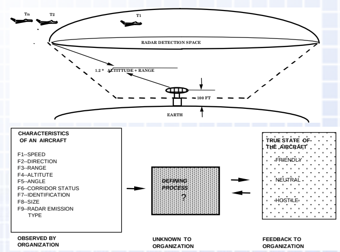
# Corp (A predecessor to ORGAHEAD)




Used a subset of capabilities that matched the human experiment




Task




CASOS

June 13, 2006  15

Carnegie Mellon

# Matched Analysis

## Simulation

- Vary organizational design
- Vary task environment
- Measure performance as accuracy
- Monte Carlo 19683 cases
- Estimate of performance on average

## Corporate Data

- Vary organizational design
- Vary task environment
- Measure performance as actual/potential severity
- General performance
- 69 cases, technological disasters

**Matched Set**

Predict performance

What if analysis: if organization did/did not shift what would be impact

CASOS

June 13, 2006  16

Carnegie Mellon

# Archival Match

**Prediction**

| Training | Performance in General | Performance During Crisis |
|---|---|---|
| Experiential | 2.10(21,0.18) | 2.38(21,0.11) |
| Operational | 1.83(48,0.10) | 1.42(48,0.07) |

**Observation**

| Training | Performance in General | Performance During Crisis |
|---|---|---|
| Experiential | 1.86(21,0.17) | 2.38(21,0.13) |
| Operational | 1.83(48,0.09) | 1.46(48,0.08) |

*Note: Number of cases and standard errors are in parentheses*

CASOS June 13, 2006  17

Carnegie Mellon

# Sub-Optimalities and Performance

**Prediction**

| Training | 1 | 2 | 3 |
|---|---|---|---|
| Experiential | 2.67(9,0.17) | 2.18(11,0.12) | 2.00(1,0.00) |
| Operational | 1.21(24,0.15) | 1.68(22,0.10) | 1.00(2,0.50) |

**Observation**

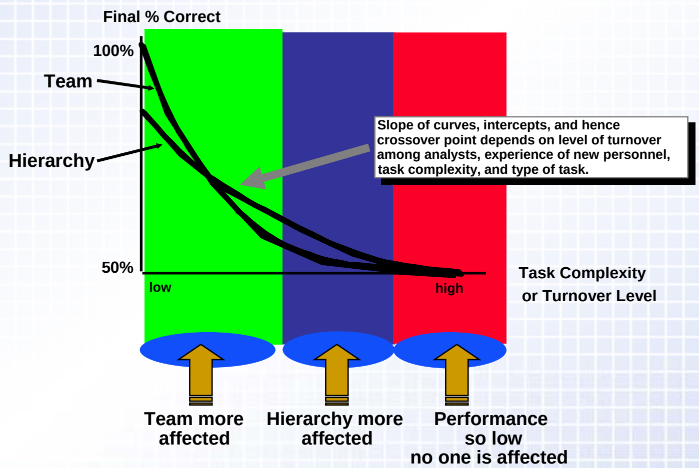
| Training | 1 | 2 | 3 |
|---|---|---|---|
| Experiential | 2.56(9,0.18) | 2.27(11,0.20) | 2.00(1,0.00) |
| Operational | 1.42(24,0.15) | 1.50(22,0.11) | 1.50(2,0.50) |

*Note: Number of cases, and standard errors are in parentheses*

CASOS June 13, 2006  18

Carnegie Mellon

## Hierarchies More Robust




CASOS

June 13, 2006



19

Carnegie Mellon

## Reality: Teams Better

| | TEAM exp | TEAM sop | HIERARCHY exp | HIERARCHY sop | SEGREG exp | SEGREG sop | NON-SEG exp | NON-SEG sop |
|---|---|---|---|---|---|---|---|---|
| Simulated | 3.00 | 1.50 | 2.35 | 1.41 | 2.30 | 1.40 | 2.45 | 1.60 |
| Human | 3.00 | 1.50 | 2.35 | 1.46 | 2.10 | 1.42 | 2.64 | 1.80 |
| | (1) | (4) | (20) | (44) | (10) | (43) | (11) | (5) |

CASOS

June 13, 2006



20

Carnegie Mellon

# Simple Validation

- Simple techniques for seeing of model results are reasonable.
- Techniques to demonstrate validity:
  - Are there stereotypical (stylized) facts about the problem that this model generates; E.G., Models of organizational evolution should predict liability of newness.
  - Are there behaviors that any model of this ilk should generate; E.G., All diffusion models should generate an s-shaped adoption curve, all neural networks should take a long time to train.
- These are non-surprising findings but if model can't generate them then it is not valid.

CASOS

Carnegie Mellon

# Docking – Model – to Model

Docking involves aligning computational models to each other

may require programming (adding additional IO routines)

may require running specific virtual experiments

Have available model description and code

Have available results from virtual experiments

Docking aids in scientific cumulation

CASOS

# Model –to-Model Analysis

- Computational models grounded in same theory may vary in implementation
  - Data representation
  - Algorithms
  - Assumptions
- Differences in implementation may significantly affect results
- Need to map out sources and effects of differences

# Docking

Definition

*"Docking (or alignment) is needed to determine whether two models can produce the same results, which in turn is the basis for critical experiments and for the test of whether one model can subsume another."*

***>>Aligning Simulation Models: A Case Study and Results (Axtell et al. 1996)***

## Model Equivalence

**Relational equivalence**

**two models produce the same internal relationship among their results**

**e.g. both show that a is a quadratic over time**

**Distributional equivalence**

**two models produce distributions of results that cannot be distinguished statistically**

**Numerical equivalence**

**two models produce numerically identical results** *not expected for stochastic models*







## More than Equivalence

- Sensitivity analysis on model features
  - Uncover the extent to which representational and operational features affect the results of a model.
- Basis for interfacing the two models
  - Can we use the output of one model as input into the other?
  - Can the models be combined into a meta-model?
- Model validity

Carnegie Mellon

## Factors Facilitating Docking

- code written to maximize model generality and ease of change
- extremely simple models
- recency of models (is data and code easily accessible)
- underlying framework compatibility - e.g., both agent oriented
- model written with docking intent
- utilizations of standard modules - e.g. random number generators
- not necessary to dock entire model

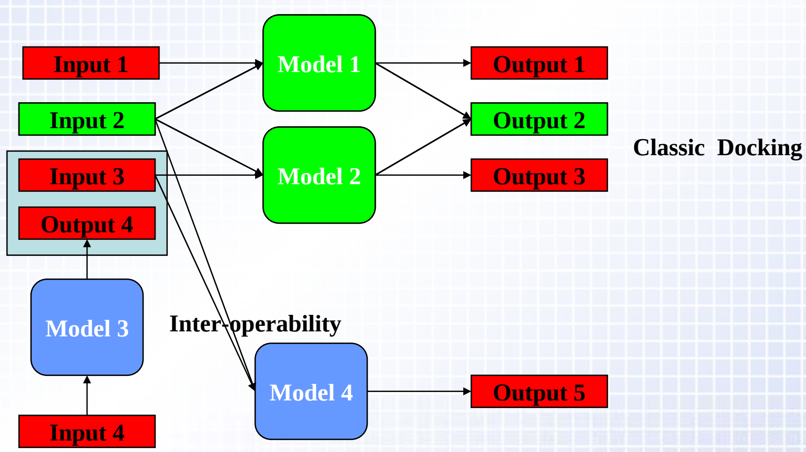
CASOS





Carnegie Mellon

## Types of Docking




CASOS

Carnegie Mellon

# Illustration 1

- Sugarscape
  - Artificial life
- Cooperation
- Rewrote smaller model
- Added data capturing routines
- Demonstrated similar results at relational level, sort-of at distributional level





Carnegie Mellon

# Illustration 2

- SimVision (Vite' VDT)
- ORGAHEAD
- Added data capturing devices
- Demonstrated relational similarity
- To be presented this year at CASOS 2002

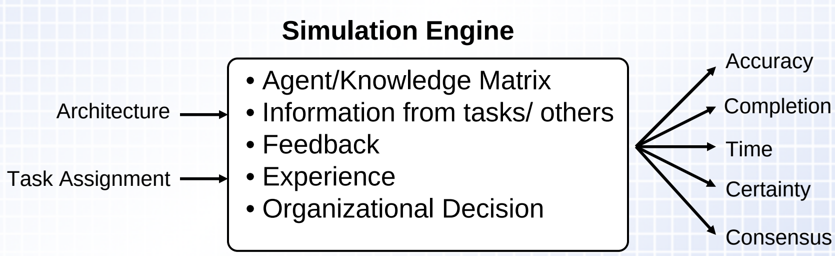
# Orgahead – SimVision docking

- Docking experiment objectives
- Model three cases (A04,A06, A16) in Orgahead, SimVision
- Analyze projects by two methods: Orgahead, SimVision
- Compare similarities/differences in methods & analysis results
- Orgahead uses “meta” matrices to define inputs
- Precedence (task x task) - Same for A06, A16
- Assignment of actors to tasks (actor x task) - Different for A06, A16, A04
- Resources, i.e., available skills (people x skills) - Different for A04,A06, A16
- Needs, i.e., required skills (tasks x skills) - Same for A06, A16
- Authority: actor reporting - Different for A06, A04,A16

 31

# ORGAHEAD 2002

## Simulation Engine for Organizational Adaptation

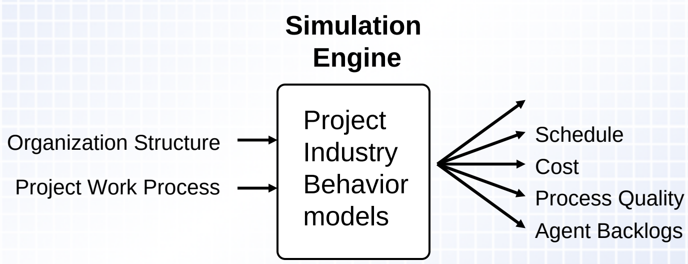



 32

Carnegie Mellon

# SimVision

## Simulation Engine for Organizational Design and Fast-track Project Execution




CASOS

Carnegie Mellon

# Orgahead – SimVision features

| | Orgahead | SimVision |
|---|---|---|
| Tasks | ✓ | ✓ |
| Task duration | | ✓ Soon |
| Task required skills | [1-n] | [1] Assume first required |
| Task precedence | | ✓ |
| Actor responsibility | [0-n] | [1] Assume first specified is primary; others secondary |
| Actor provides feedback | ✓ | |
| Task rework | | ✓ (N/A for this experiment) |
| Coordination | Actor - actor | Task – Task (Not included in 7.11 analysis) |

CASOS

Carnegie Mellon

# Encoding the Data

| | ORGAHEAD 2002 | SimVision |
|---|---|---|
| Authority Structure | Yes | Yes |
| Resource Requirements | 1 to many | Assume 1st required |
| Resource Assignment | 1 to many | Assume 1st required |
| Tasks | Binary bit string | Symbolic |
| Task Precedence | Simulated | Yes |
| Task Assignment | 1 to many | Assume 1st specified |

CASOS

June 13, 2006  35

Carnegie Mellon

# Orgahead – SimVision Task – actor assignments

Case A06

Actors (rows) by tasks (columns)

Yellow cells show tasks assigned by default to the actor with the best skill match.

Blue cells show tasks and their assigned secondary actors

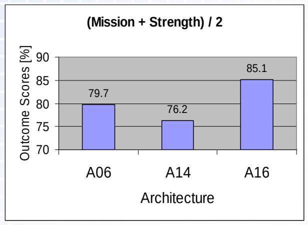
Task actor assignments A06 (actor x task)

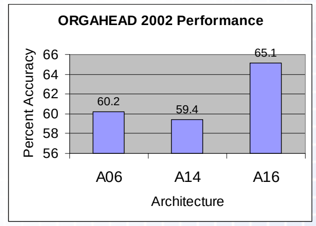
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 |
| 6 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |

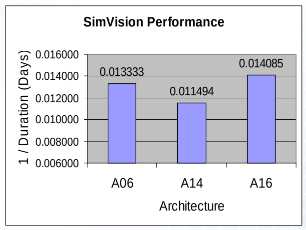
CASOS

June 13, 2006  36

# Results: Accuracy vs. Duration



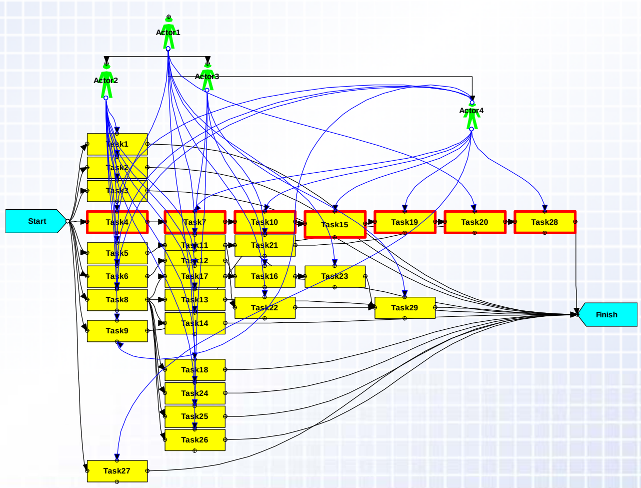




| | |
|---|---|
| SimVision | A14 < A06 < A16 |
| ORGAHEAD | A14 ≤ A06 < A16 |
| Real Data | A14 < A06 < A16 |

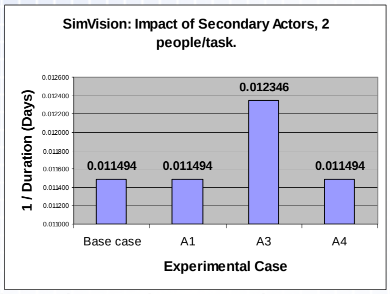
  

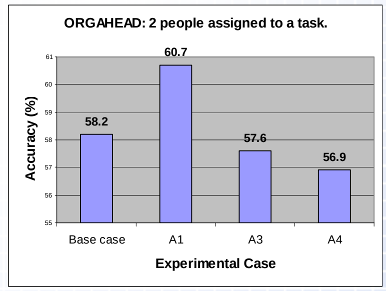
# Results
## Effects of Secondary Actors



# Results
## Effects of Secondary Actors




- What other tasks secondary actors are assigned to matter.
- Who is assigned the primary actor can significantly alter the results.





# Illustration 3

- ELM
- Plural-Soar

Carnegie Mellon

# Comparison of Models

Carley, Kjaer-Hansen, Prietula & Newell 1991

| | Plural-Soar | ELM |
|---|---|---|
| **Perception and Action** | | |
| Perceives the environment | ❍ | ❍ |
| Physically manipulates objects | ❍ | |
| Moves self to different locations | ❍ | |
| **Memory** | | |
| Location | ❍ | ❍ |
| People | | ❍ |
| Task | | ❍ |
| **Instruction** | | |
| Can be incomplete | ❍ | ❍ |
| **Task Analysis** | | |
| Decomposes task | | |
| Coordinates subtasks for self to do | | |
| **Communication Skills** | | |
| Asks questions/Provides answers | ❍ | ❍ |
| Gives commands/Receives commands | ❍ | ❍ |
| Talks to a single individuals/Talks to a group | ❍ | ❍ |
| **Social Analysis** | | |
| Models of other agents | | ❍ |
| Model of organization | | |

CASOS

June 13, 2006  41

Carnegie Mellon

# Comments on Docking

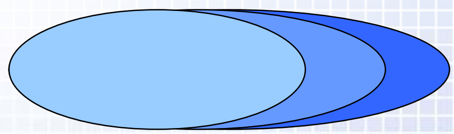
- Docking models may force non-trivial data encoding decisions
- Data encoding decisions can significantly affect results
- The docking process should include sensitivity analyses on model features

CASOS

June 13, 2006  42

## Predictive Docking

- General theory of what features are necessary to get what outcomes
- Generate a mapping of features to outcomes such that as you increase in features you increase in outcomes and later features retain the properties of earlier ones
- Locate model or models on this map by its features
- Models in the same map location are directly dockable
- Models earlier in the mapping are contained in those later in the mapping





Cognitive Architecture — Increasingly Limited Capabilities

Knowledge — Increasingly Rich Situation

| | Nonsocial Task | Multiple Agents | Real Interaction | Social Structural | Social Goals | Cultural Historical |
|---|---|---|---|---|---|---|
| Omniscient Agent | goal directed<br>produces goods<br>uses tools<br>uses language | models of others<br>turn taking<br>exchange | face-to-face | class differences | organizational goals | historically situated |
| Rational Agent | reasons<br>acquires<br>learns | learns from others<br>education<br>negotiation | mis-communication | promotion<br>social mobility | competition<br>cooperation<br>social cognition | emergent norms<br>*Cultural Transmission* |
| Boundedly Rational Agent | satisfices<br>task planning<br>adaptation | group making | social planning<br>coercion<br>priority disputes | altruism<br>uses networks for information<br>boundary spanners<br>*Garbage Can Model*<br>*Sugarscape, AAIS* | delays gratification<br>moral obligation<br>*VDT*<br>*TAEMS* | gate keeping<br>role emergence<br>*CORP, HITOP-A, ACTION, ORGAHEAD, Organizational Consultant* |
| Cognitive Agent | compulsiveness<br>lack of awareness<br>multi-tasking | group think | spontaneous exchange<br>social interactions | automatic response to status cues | group conflict<br>power struggles<br>*TAC Air Soar*<br>*Plural-Soar* | develop language<br>institutional change |
| Emotional Cognitive Agent | habituation<br>variable performance | protesting<br>trust | play<br>rapid emotional responce<br>cons | campaining | team player | norm maintenance<br>ritual maintenace<br>advertising<br>**MODEL SOCIAL AGENT** |

Carnegie Mellon

# Statistical Comparison Approach

- Statistical comparison of the predictive ability of multiple models
  - The predictions of each model for a set of data are generated
  - These predictions are compared with non-computational data
- In between harmonization and docking
- Involves contrast of multiple models, some of which might be computational
- Assesses theoretical adequacy of model
- Assesses points of comparability between models in terms of predictions

CASOS

June 13, 2006  45

Carnegie Mellon

# Illustration

- Data:
  - Kapferer study of workers in tailor shop in Zambia
  - Network data at time 1 and 2
  - Socio-demographic, and experience information
- Approach:
  - For each model, given the data at time 1, generate model's predictions for time 2
  - Given predicted time 2 and real time 1 calculate the number of ties that: stayed, dropped, added, remained no-tie
  - Contrast this with number of ties for real time 2 given real time 1 that: stayed, dropped, added, remained no-tie
  - Look at overall R2

locate areas of relative effectiveness

CASOS

June 13, 2006  46

Carnegie Mellon

# Change in Interaction

| MODEL | R | never | continue | begin | stop | Total |
|---|---|---|---|---|---|---|
| Random | .000 | 572<br>67% | 74<br>33% | 91<br>33% | 85<br>67% | 822<br>56% |
| Kapferer's Exchange | .225 | 632<br>74% | 168<br>76% | 108<br>39% | 4<br>3% | 912<br>62% |
| Heiderian Balance | .339 | 224<br>26% | 222<br>100% | 242<br>88% | 0<br>0% | 688<br>46% |
| | | | | | | |
| CONSTRUCT | .422 | 662<br>77% | 215<br>97% | 140<br>51% | 1<br>1% | 1018<br>69% |

CASOS

June 13, 2006  47

Carnegie Mellon

# Synopsis

- Many different types of validation and docking
- Varies by level
- Varies by equivalence

- V&D need not always be done
- V&D should be appropriate to intent
- V&D requires teams
- V&D often best not done by modeler

CASOS

June 13, 2006  48

Carnegie Mellon

# Guidelines & Critiques

Center for Computational Analysis of Social and Organizational Systems

http://www.casos.cs.cmu.edu/

Carnegie Mellon

## Guidelines

- KISS
- Keep the key experiment in sight
- No theory is sufficient
- Take a building block approach
- Be willing to start over
- Keep virtual field notes, a lab notebook
- Minimize the number of parameters
- Don't hardwire parameters
- Don't hardwire input data
- if you must use fixed parameters have a low, medium, high option
- if you don't run a virtual experiment for this parameter then monte-carlo across it (randomly set its value for each run)

June 13, 2006  50

Carnegie Mellon

# Coding Levels

| 25% | 25% | 50% |
| --- | --- | --- |
| Input | Theory | Output |

**Capture data in lots of ways**

**this is important for debugging, testing, and determining what to change**

**Trace**
**Over time values**
**Final Values**
**Types of Changes**
**Mean, Variance, N ,**
**$\Sigma$ X, $\Sigma$ X2**

CASOS



Carnegie Mellon

# Visibility & Linking

- Make all parameter values external to the system
  - Enables changing values at will
- Enable tracing
- Make all output in CSV format

CASOS

Carnegie Mellon

# Entity Segregation

| Entity | I/O | Environment |
|---|---|---|
| Person | get and send message | other persons |
| | perceive and affect | organization |
| | | external world |
| Organization | perceive and affect | internal agents |
| | form and break linkages | other organizations |
| | | external world |

Only allow interaction through specific I/O interface

Increases speed, admits model expansion, higher validity, avoids errors and certain pathologic behavior, aids timing

**Keep entities as distinct objects with their own I/O for interacting with other entities and the environment**

CASOS

June 13, 2006  53

Carnegie Mellon

# Appropriate Model Critiques

- Are the following model components described?
  - Inputs
  - Outputs
  - Internal mechanisms or processes
  - Initial conditions
  - Parameters
  - Boundary conditions
  - Limits to the model
- Is the model well specified; I.E., Could another researcher adapt this model or build on it based on this description?
- How does this model relate to other computational models?
  - Is the literature cited
  - Are comparisons and distinctions drawn

CASOS

June 13, 2006  54

# Appropriate Virtual Experiment Critiques

- Are the following aspects of the virtual experiment described?
  - Source of data
  - Variables
  - Possible biases
  - Limits of data set
- How well are analysis methods described?
- Should other analysis methods been employed?

# Inappropriate Critiques

- Model should not be presented until it has been validated (that would require a second research paper)
- Model alterations that would require significant reprogramming (multiple months) and substantial time running virtual experiments (multiple months) (that would require a second research paper)
- For validation studies, only present the empirical data and not the computational data (that misses the point of the paper)
- Present everything as a flow chart (not possible with some-types of concurrent systems)
- Code must be provided (not possible under some contracts and code may be too extensive)
- R2 for linear models are too low for the analysis of the virtual experiments (R2 is irrelevant as it is known that the model is non-linear)