How To Make Users Happy

And avoid annoying users

User Happiness?

\[ H_u = f(\text{Performance}) \]

User Happiness?

\[ H_u = f(\text{Performance, Trust}) \]
User Happiness!

$$H_u = \ell (E_{\text{Assistant}}, E_{\text{Negative}}, E_{\text{Positive}}, E_{\text{Value}}, E_{\text{User}})$$

- E_{\text{Corrected}}
- E_{\text{By-hand}}
- E_{\text{Cost}}
- E_{\text{Avoided}}
- E_{\text{Correct-difficulty}}
- E_{\text{Sensible}}
- W_{\text{Quality}}
- W_{\text{Commitment}}
- T_{\text{By-hand}}
- T_{\text{By-Hand-start-up}}
- T_{\text{By-Hand-per-unit}}
- T_{\text{Assistant}}
- T_{\text{Training-start-up}}
- T_{\text{Assistant-per-unit}}
- T_{\text{Interaction-per-unit}}
- T_{\text{Monitoring}}
- T_{\text{Correcting}}
- T_{\text{Responsiveness}}
- T_{\text{System-Training}}
- T_{\text{User-training}}
- T_{\text{Average-for-each-correction}}
- T_{\text{Error-rate}}
- N_{\text{units}}
- P_{\text{Pleasantness}}
- U_{\text{Perceive}}
- U_{\text{Why}}
- U_{\text{Provenance}}
- U_{\text{Predictability}}
- T_{\text{Assistant-interfere}}
- T_{\text{Screen-space}}
- I_{\text{Cognitive}}
- I_{\text{Appropriate-Time}}
- C_{\text{Autonomy}}
- C_{\text{Correcting}}
- S_{\text{Sensible-Actions}}
- S_{\text{User-models}}
- S_{\text{Learning}}
- R_{\text{Social-Presence}}
- D_{\text{Hand}}
- V_{\text{Importance}}$$

Why Happiness?

- Focus on assistants that take on tasks which the users could do themselves
- Assistants are supposed to be helpful
- If not, users can turn off the assistants
  - Optional
  - Assume: cannot require users to use assistant or to provide feedback
- So only used if user finds it:
  - Useful
  - Trustable
  - Usable

Adjustable Autonomy

- Assistant does it all; completely autonomous
- User monitors actions of assistant (confirmation of assistant’s actions)
- Assistant helps user do actions (or user tells agent how to do the actions)
- Assistant tells users where actions might be done
- User does all actions; direct manipulation

Key Factors

- Correctness
  - Errors
- Speed
  - Time to use system with the assistant
- Pleasantness
- Utility
Measures for Correctness

- Can measure % correct on corpus
- Or measure in field deployment
  - Often performance is much worse
- Also important is:
  - Overhead of monitoring for correctness
  - Time for correction
- If Assistant can be wrong, user might need to check each action
- When is wrong, need to:
  - Notice is wrong
  - Fix the error
- How long does this take compared to just doing it?
- But doing it by hand might have errors too!

Correctness

- How many errors does the assistant make?
  - $E_{\text{positive}}$: False negatives: missed opportunities to help ("coverage")
    - Just silent when might do something
  - $E_{\text{false}}$: False positives: incorrectly offered to help ("precision")
  - $E_{\text{value}}$: Wrong values: partially correct, but with inaccurate parameters
- Total errors left in the results
  - $E_{\text{user}}$: User’s errors also involved
  - $E_{\text{corrected}}$: User might catch errors and fix them
- $E_{\text{assistant}} = E_{\text{user}} + E_{\text{positive}} + E_{\text{value}} - E_{\text{corrected}}$
- By-hand: But compare to errors when no assistant
- Error rate may change over time, as the assistant learns

Examples

- Radar VIO (Virtual Information Officer) helps fill in form fields from emails

VIO Error Rate

- With VIO: Overall decrease in time by 17% ($p < .001$)
- Overall error rate (all users)
  - $E_{\text{assistant}}$: 15 (total errors left in result) vs.
  - $E_{\text{by-hand}}$: 12 ($n.s.$)
- Per user error rates (20 users):
  - $E_{\text{negative}}$: 12 (missed extracting values)
  - $E_{\text{positive}}$: 0
  - $E_{\text{value}}$: 1
- VIO strong biased away from incorrect guesses, so prefers not to say anything (favors $E_{\text{negative}}$)
Example: Citrine Errors

- Interprets addresses in copied text
- Copy-and-paste by hand for people’s addresses took more time even including fixing errors, compared to using the Citrine assistant
- When by hand: left more errors in result


Old Example

- Peridot (1985); confirm by question and answer
- Low consequence of errors
- Users generally just said “Yes” without understanding the question
- Assumed computer knew better than they did
- So can’t necessarily trust user’s feedback


Consequences of Errors

- Not just a factor of the time for errors
- Other factors:
  - $E_{Cost}$ Cost (seriousness) of making an error
    - Probably a key factor in user’s acceptance and happiness
  - Aircraft auto-pilot vs. filling in addresses for a contact
  - Likelihood of making an error by hand compared to by the assistant (error avoidance):
    - $E_{Avoided} = E_{By-hand} - E_{Assistant}$
  - $E_{Apparentness}$ Likelihood of noticing an error
  - $E_{Correct-difficulty}$ Ease of correction of the error
    - Likelihood of being able to correct it after finding it
    - Is the right information available?

Quality of Errors

- User happiness may not only depend on frequency and severity of errors
- Henry Lieberman: Depends on whether the errors make sense
  - Predictable vs. seemingly random errors
  - Knowledge-based vs. statistical techniques
  - But often errors easier to notice if very far off
    - Example: OCR, mistakes in Citrine
- Helps users predict how to avoid errors
- $E_{Sensible}$
Quality of Work Beyond Errors

- May not be right vs. wrong
- Quality of the assistant’s work
  - Mary Shaw: Satisfactory level of work
    - E.g., Meeting transcripts
  - \( W_{\text{Quality}} \)
- Wayne Iba
  - Bad answers may inspire user to better work
    - Apprentice
  - \( W_{\text{Commitment}} \): User’s attitude and commitment affects quality

Measuring Time

- Time when performing tasks
- Can measure the time for the user without assistant, compared to with assistant
- Can include time to correct errors
  - But only those that the user notices
  - Corrected errors vs. Un-corrected errors
- Usually want the time to be faster when using the assistant
  - May be slower for 1st time, but faster if used a lot
  - Because of training, learning time, etc.
- Does not include “background” time
  - Assistant can work in parallel to user

Equations for Time

- Control condition:
  \[ T_{\text{By-hand}} = T_{\text{By-hand-start-up}} + (T_{\text{By-hand-per-unit}} \times N_{\text{units}}) \]
- Time with assistant, including errors:
  \[ T_{\text{Assistant}} = T_{\text{Training-start-up}} + (T_{\text{Assistant-per-unit}} \times N_{\text{units}}) \]
- Where:
  \[ T_{\text{Assistant-per-unit}} = T_{\text{Interaction-per-unit}} + T_{\text{Monitoring}} + T_{\text{Correcting}} + T_{\text{Responsiveness}} \]
  \[ T_{\text{Correcting}} = A_{\text{Error-rate}} \times T_{\text{Average-for-each-correction}} \]
  \[ A_{\text{Error-rate}} = \frac{E_{\text{Positive}} + E_{\text{Value}}}{T} \]

Time per Item

- \( T_{\text{Interaction-per-unit}} \): is average across items
  - Ones handled correctly by assistant lowers average time
  - If agent anticipates and does task, then small or 0
  - Should be lower than \( T_{\text{By-hand-per-unit}} \) or will never win
- \( T_{\text{Responsiveness}} \): Includes time that user has to wait for assistant
  - If agent slows down interaction
  - Also, if agent is slow, makes it look stupid
  - People don’t like to wait even if overall is faster
    - Xerox Star judged poorly even though overall faster tasks
  - Conversely, people feel fast when busy with DM
**Time and Accuracy**

- What accuracy rate is required?
  \[ T_{by\text{-}hand} \geq T_{assistant} = T_{training\text{-}start\text{-}up} + (T_{interaction\text{-}per\text{-}unit} + T_{monitoring} + (A_{error\text{-}rate} \times T_{average\text{-}for\text{-}each\text{-}correction})) \times N_{units} \]

- This formula can help determine how much accuracy is required for assistant to be worthwhile
- Can improve performance by improving UI for monitoring and correcting!
- If importance of checking is low, then user might not check any/all

---

**Issues with Training Time**

- \( T_{training\text{-}start\text{-}up} \) includes system training and user training
  \[ T_{training\text{-}start\text{-}up} = T_{system\text{-}training} + T_{user\text{-}training} \]

- \( T_{system\text{-}training} \) is explicit training requires
  - Might be labeling examples, entering rules, specifying policies & permissions, etc.
  - CALO users complain about re-training required with every new release
  - Other assistants "pre-train" on corpus or do not need training, so: \( T_{system\text{-}training} = 0 \)
  - Try to get training from what users do anyway (implicit) so no extra overhead

- \( T_{user\text{-}training} \) is time for user to learn how to use the assistant

---

**Example of Performance Measures**

- When there are repeated tasks, can measure cross-over point (N_{units})
- When there are enough tasks to overcome the overhead
  - Example, LAPIS supported "simultaneous editing"
    - Teach a pattern and edit all locations at once
      - Robert C. Miller and Brad A. Myers.

---

**Example of Performance Measures**

- Measure when enough tasks to overcome the overhead

<table>
<thead>
<tr>
<th>Task</th>
<th>Records-per-task</th>
<th>Simultaneous editing</th>
<th>Manual editing</th>
<th>Time-per-task (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>42.3s (42.3s)</td>
<td>21.0s (21.0s)</td>
<td>3.4 sec (3.4 sec)</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>115.1s (120.9s)</td>
<td>22.6s (29.8s)</td>
<td>3.6 sec (4.0 sec)</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>150.0s (147.8s)</td>
<td>21.3s (29.1s)</td>
<td>3.8 sec (4.0 sec)</td>
</tr>
</tbody>
</table>
Subjective Factors

- How much do users like the assistant?
- Can be annoying even when not doing anything
- Alternatively, might be considered positively
  - Cute, helpful, polite, ...
- $P_{\text{Pleasantness}}$

Factors of Pleasantness

- Understandability
- Interference
  - Interruptions
- User control
- Sensible Help
- Social Presence

$$P_{\text{Pleasantness}} = f(U_{\text{Perceive}}, U_{\text{Why}}, U_{\text{Provenance}}, U_{\text{Predictability}},$$
$$I_{\text{Assistant-interfere}}, I_{\text{Screen-space}}, I_{\text{Cognitive}}, I_{\text{Appropriate-Time}},$$
$$C_{\text{Autonomy}}, C_{\text{Correcting}}, S_{\text{Sensible-Actions}}, S_{\text{User-models}},$$
$$S_{\text{Learning}}, R_{\text{Social-Presence}})$$

Understandability

- Does user understand what is happening?
- Related user interface principles (Nielsen’s Heuristics):
  - Visibility of system status, Recognition rather than recall, Aesthetic and minimalist design, Help users recognize, diagnose, and recover from errors, Help and documentation
- User able to perceive what the system is doing
  - $U_{\text{Perceive}}$ Actions, states, reasons are visible
- Understand why actions are being taken (“Transparency”)
  - $U_{\text{Why}}$ Lots of work on this topic
  - $U_{\text{Why}}$ And understand the assistant’s answers
- Interacts with control
  - Not just understand why
  - Also, be able to change or fix it
  - Not do it the same way next time

Understanding Values

- Understand actions assistant does
- Also important: understanding values
  - Where the values come from
  - Conley & McGuinness: Provenance and Credibility of values
  - $U_{\text{Provenance}}$
Explaining Why

- Crystal – my system to explain “why” for assistants in complex applications like Microsoft Word
- Doesn’t just explain why, but brings up the dialog boxes to let user change it

Crystal Answers

Predictability

- Can the user predict what the agent will do?
- Related user interface principles:
  - Consistency, Visibility of system status, Match between system and the real world
- Predictable <-> understandability
  - Understand future actions
  - Not just what it has already done
- UPredictability

Interference

- Assistant-interfere How much does the assistant interfere with other tasks?
  - Can make user less effective on unrelated tasks
  - Screen-space Screen space for the assistant
  - Compare Clippy vs. squiggly underlines
  - Towel’s To-Do list window; TamaCoach’s GUI
  - Really big explanations (Crystal, CALO, etc.)
  - Radar repeats email with the assistant’s interpretation
Interference

- I_{Cognitive} Cognitive overhead of monitoring assistant
  - Attention taken away from other tasks
  - Example: Meeting Rapporteur mentions checking/correcting assistant’s notes compared to participating in meeting
    - Vs. taking notes by hand
- T_{Monitoring} Time overhead already included

Interruptions

- Interruptions interfere
- Can be annoying
- But may be necessary
  - E.g., RoboCare notification for medicine
- Some systems trying to predict appropriate times to interrupt
- Decisions:
  - Whether to interrupt
    - Vs. perform autonomously or not assist at all
  - How to ask the question (understandability)
  - When to interrupt

Interruptions, example

- Radar Attention Manager
  - Dan Siewiorek & Asim Smailagic

Radar Attention Manager

- Subject rating of interruption annoyance (1 Low, 10 High)
  - Subtask boundaries worse than random; tasks boundaries better
- Good success at predicting when interruptible

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>True positives</th>
<th>True negatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSVM</td>
<td>85.35 %</td>
<td>82.69 %</td>
<td>88.90 %</td>
</tr>
<tr>
<td>SVM</td>
<td>81.24 %</td>
<td>80.77 %</td>
<td>81.71 %</td>
</tr>
<tr>
<td>LDA</td>
<td>75.47 %</td>
<td>69.23 %</td>
<td>81.71 %</td>
</tr>
<tr>
<td>MNN</td>
<td>66.74 %</td>
<td>55.77 %</td>
<td>77.71 %</td>
</tr>
</tbody>
</table>
User Control

- Ability of user to control the assistant
- Related user interface principles:
  - User control and freedom, Error prevention, Flexibility and efficiency of use, Help users recover from errors
- Control autonomy, related to assistant overhead: \( T_{\text{Training-start-up}} \)
- Correcting difficulty of fixing results of errors
  - Related to \( T_{\text{Monitoring}} + T_{\text{Correcting}} \)
  - Also possibly mental difficulty of doing this process
    - Not just time

Sensible Assistance

- Sensible-Actions, whether the proposed actions make sense
- Related to \( E_{\text{sensible}} \) for errors
- “Don’t Be Stupid”
  - Not keep asking the same thing over and over
- Requires:
  - User-models User modeling, so answers are appropriate
  - Learning, so answers change

Social Presence

- Users may relate better to animated agents
- But “Uncanny Valley”
  - Theory that if agent looks and behaves too much like a person, but not quite, then much worse
  - Increased by movement
  - Linked to zombies and death


Summative Measures

- Utility
- Trust
- Performance
Utility

- How much Value is what the assistant does?
- How much Work (effort) does it take?
- \[ \text{UTILITY (usefulness)} = \frac{\text{Value}}{\text{Work}} \]
- \[ \text{Value} = F(D_{\text{hand}}, V_{\text{importance}}, E_{\text{avoided}}) \]
  - \[ D_{\text{hand}} \text{ How difficult would the task be to do by hand?} \]
  - \[ V_{\text{importance}} \text{ How important is it to do the task?} \]
  - \[ E_{\text{avoided}} \text{ Errors avoided } \]
- Work (effort)
  - Partially T_{\text{assistant}}
  - Maybe include mental workload, etc.

Trust

- What are the factors that go into Trust?
- (Lots of good talks on this topic Tuesday)
- All of the Error metrics
  - Number, cost of errors
  - Ease, likelihood of correcting errors
  - False positives (false alarms) particularly damaging
- Understandability
  - Visibility of what doing
  - Why doing it
- Maybe all the factors?
  - Not just explanations

Others?

- Wayne Iba lists:
  - Competence
  - Attention
  - Anticipation
  - Persistence
  - Deference
  - Integrity
  - Picking appropriate task to automate
- Christopher Miller, et. al. lists other risks
  - Lack of situation & system awareness
  - Increase in user’s mode errors
  - Too much trust can also be bad
  - Automation causes increased workload
- Nadine Richard & Seiji Yamada lists “fun factor”
- Are these covered by the factors?

Issue: Converting Do’ers to Managers

- Converting from Direct Manipulation to managing assistants
  - But managing is hard
    - The most valuable jobs are managers:
      - Terry J. Semel, CEO Yahoo, $238.5 mil
      - Barry Diller, CEO IAC, $156.2 mil
      - Tiger Woods $80.3 mil
      - Tom Cruise, $31 million
  - People have to learn how to effectively use human helpers
  - Also, user may know “right” answer only by constructing it
  - Need to “directly manipulate” to investigate the answer
  - Don’t assume that converting a task to a managerial one will inherently make it easier!
Perceived Costs and Performance

  In Proceedings of the IEEE Symposia on Human-Centric Computing Languages and
  Environments, pp. 2-10.
- Given a choice, users evaluate cost-benefit:
  - Investment – learning, etc. to be ready to do task
  - Cost – to do the desired task
  - Pay-off – reduced future cost
  - Risk – probability that no future pay-off will result
  - Decision cost – cost of making this decision
- Users can’t know real values, so guess, based on
  experience, personal style, etc.
  - Easier to estimate the costs to doing task manually
  - Hard to estimate costs and risks of using assistant

Perceived Costs and Benefits

- People overrate errors, under-perceive time saved
  - Strongly prefer not to learn something new
  - Strongly prefer to avoid risk
- People don’t necessarily make rational decisions
- User interface can influence perceptions of costs & benefits
  - E.g., Incremental, small steps
- Why there might be a discontinuity in $H_u = f(...)$
  - A little better performance of assistant results in
    disproportionate gains in $H_u$

Perceived Costs When Changing

- Particularly difficult with systems that learn
- Past performance may not be a good indicator of future performance
- Need some way to indicate what learned
  - Hopefully more fluid and effective than clippy
  - Continuous instead of binary?

Usability Methods

- Conventional Usability Methods work for
  Intelligent Assistants
  - Contextual inquiry
  - Involving designers in the design process
  - Paper-prototyping
  - Wizard-of-Oz prototyping
  - Heuristic analysis
  - Think-aloud user studies
  - Etc.
- Can measure many of the values in A vs. B experiments
  - E.g., compared to the non-assisted version
  - Not appropriate to say “User can easily...” without data
Example

- Improved Radar’s task manager through iterative design with user studies
- Users didn’t understand “Confidence”, “Phase” vs. “Importance”

Summary: User Happiness?

- $H_U = f (\ldots)$
- Don’t yet know all the factors
- Certainly don’t know the function
- But ones that we do know should be measured and optimized
  - Existing HCI methods are effective
  - Worthwhile goals to investigate to get assistants that are useful, usable, & pleasant

Acknowledgements

- Thanks to many people who helped with this talk
  - Polo Chau, Andrew Faulring, Michael Freed, Sara Kiesler, Andy Ko, Henry Lieberman, Chris Scaffidi, Dan Siewiorek, Rich Simpson, Stephen Smith, Aaron Steinfeld, Anthony Tomasic, Neil Yorke-Smith, John Zimmerman

- Funded by DARPA through RADAR

User Happiness!

$H_U = f(E^{\text{Assistant}}, E^{\text{Negative}}, E^{\text{Positive}}, E^{\text{Value}}, E^{\text{User}}, E^{\text{Corrected}}, E^{\text{By-hand}}, E^{\text{Cost}}, E^{\text{Avoided}}, E^{\text{Apparentness}}, E^{\text{Correct-difficulty}}, E^{\text{Sensible}}, W^{\text{Quality}}, W^{\text{Commitment}}, T^{\text{By-hand}}, T^{\text{By-Hand-start-up}}, T^{\text{By-Hand-per-unit}}, T^{\text{Assistant}}, T^{\text{Training-start-up}}, T^{\text{Assistance-per-unit}}, T^{\text{Interaction-per-unit}}, T^{\text{Monitoring}}, T^{\text{Correcting}}, T^{\text{Responsiveness}}, T^{\text{System-Training}}, T^{\text{User-training}}, T^{\text{Average-for-each-correction}}, T^{\text{Error-rate}}, T^{\text{N-units}}, P^{\text{Pleasantness}}, P^{\text{Perceive}}, P^{\text{Why}}, P^{\text{Provenance}}, U^{\text{Predictability}}, U^{\text{Assistant-interferes}}, U^{\text{Screen-space}}, I^{\text{Cognitive}}, I^{\text{Appropriate-Time}}, I^{\text{Autonomy}}, I^{\text{Correcting}}, S^{\text{Sensible-actions}}, S^{\text{User-models}}, S^{\text{Learning}}, R^{\text{Social-Presence}}, D^{\text{Hand}}, V^{\text{Importance}})$
Issues Brought Up During Discussion

- Probably calling the top-level measure "Happiness" is incorrect, since users aren’t good at perceiving effectiveness.
  - What would be a better term for all factors together?
- What about Intelligent Tutors?
  - Need new factors for User’s learning, user’s motivation
  - The comparison is tutoring by a person
- What about when using Assistant is required, e.g. for safety, by policy?
- What about Mixed Initiative?
  - Are there new factors?
- What are the higher-level, summative factors?
  - TAssistant vs TBy-hand; EAssistant vs. EUser; Pleasantness is harder