

LET'S DISCUSS: LEARNING METHODS FOR DIALOGUE
NIPS 2016 WORKSHOP, December 10th, 2016, Barcelona, Spain

Awkward Silence? The Evaluation of Social Dialogue Systems

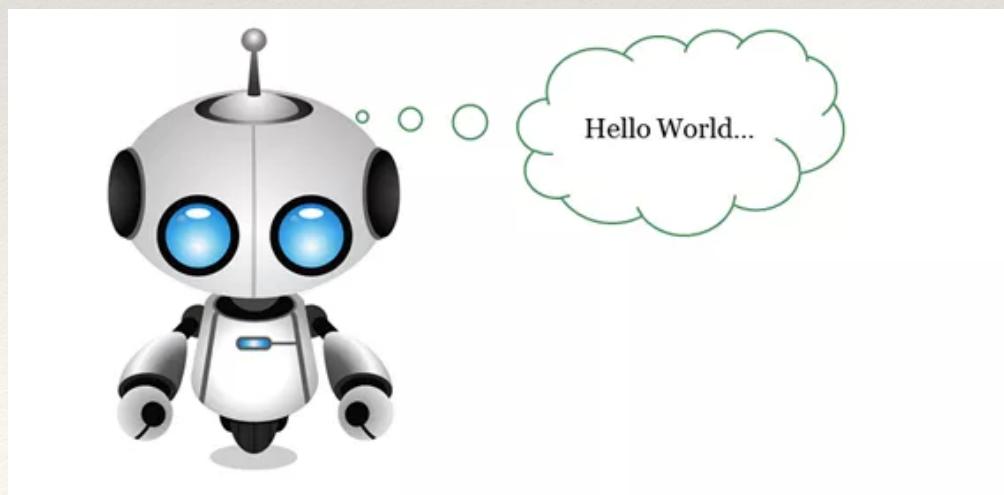
Helen Hastie
Heriot-Watt University,
Edinburgh

Overview

- ❖ What are Social Dialogue Systems (aka Social Chatbots) and why are they so tricky to evaluate?
- ❖ Survey of current evaluation metrics
- ❖ What can we learn from other disciplines?
 - ❖ gaming, HCI, psychology, robotics and cognitive theory
- ❖ Ethics
- ❖ Conclusion and summary

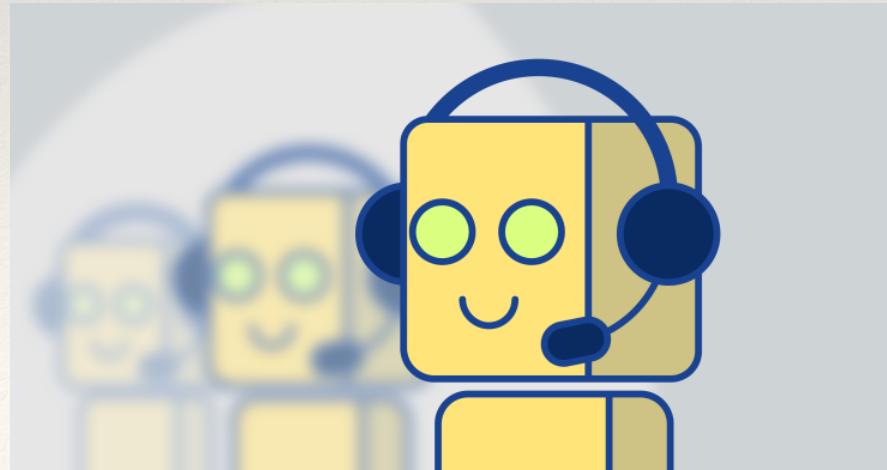
What is it?

- ❖ A Social Dialogue System / Chatbot is designed to engage humans on topics of interest,
- ❖ generate interesting, coherent, meaningful responses,
- ❖ carry on as long as possible,
- ❖ be a companion.



What's the point of them?

- ❖ Provides relevant information (What do I need to know?)
- ❖ Supports decision making (What should I do?)
- ❖ Facilitates action (How do I do something?)
- ❖ Keeps you company (I'm lonely I want to talk)



It's good to talk

- ❖ Interactive system as a companion (Wilks, 2006)



A system by any other name

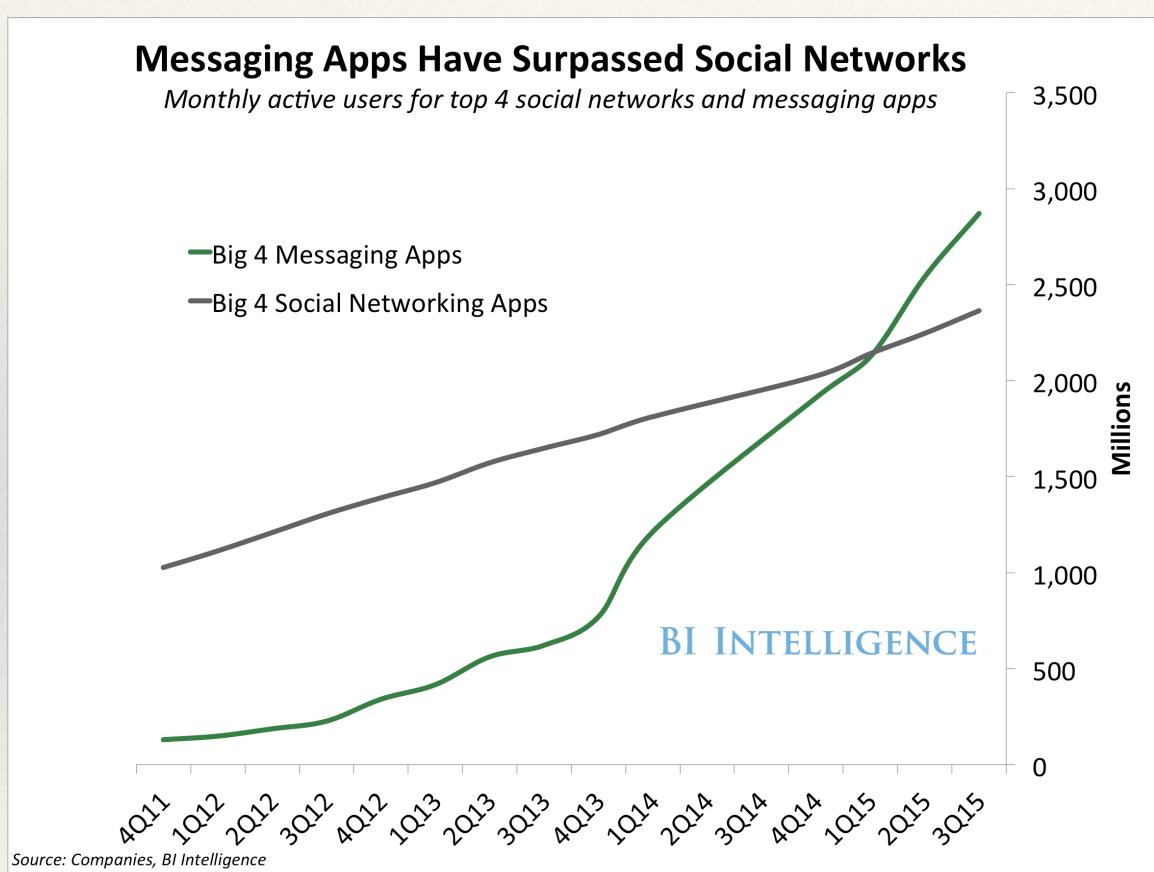
- ❖ (Neural) Response generation
- ❖ Chatterbot
- ❖ Unsupervised dialogue systems
- ❖ Dialogue chatbot
- ❖ Interactive CA
- ❖ Chatbot
- ❖ End-to-end CA
- ❖ AI bot
- ❖ Response Retrieval
- ❖ Socialbot
- ❖ Example-based dialogue systems
- ❖ Social Media Bot
- ❖ Next Utterance Classification
- ❖ Social chatbot



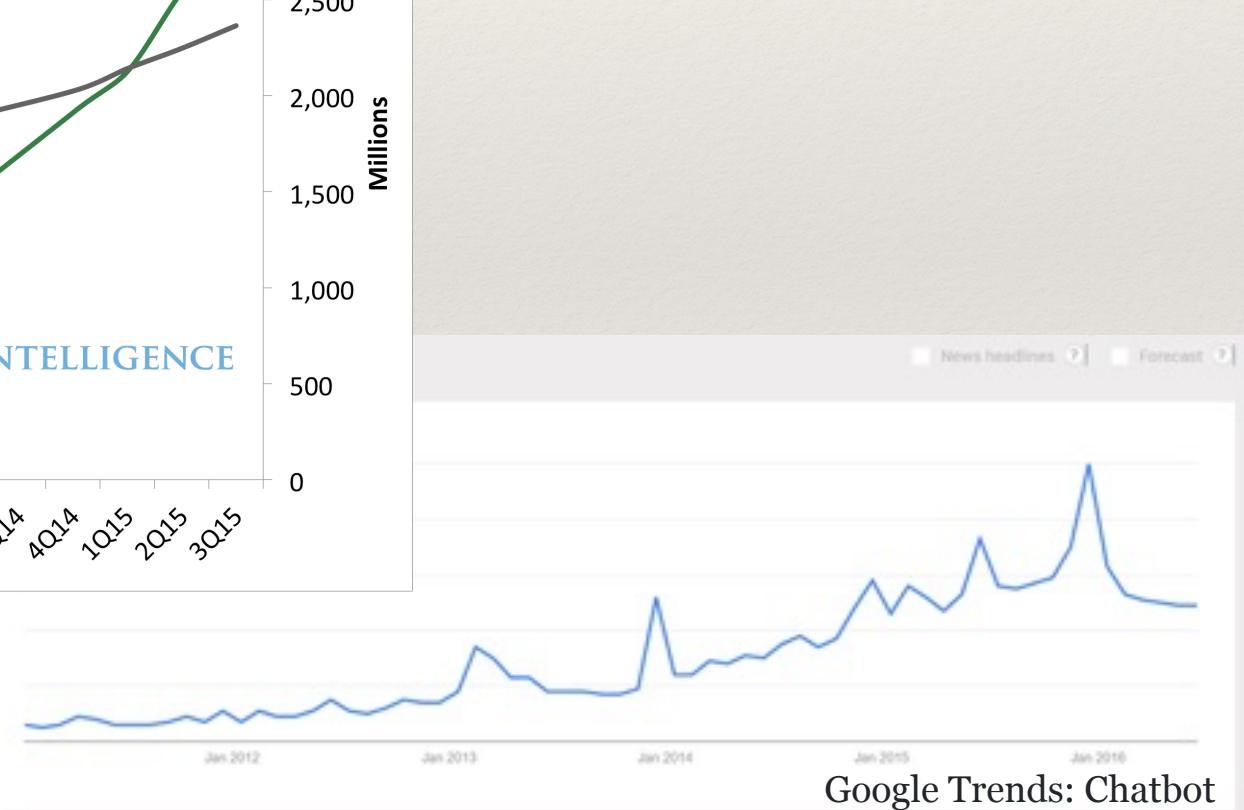
Why all the fuss?

Messaging Apps Have Surpassed Social Networks

Monthly active users for top 4 social networks and messaging apps



Bots are the new apps



The Alexa Prize

\$2.5 Million to Advance Conversational Artificial Intelligence

September 2016 – November 2017



Challenge: build a socialbot that can converse coherently and engagingly with humans on popular topics for 20 minutes.

Types of systems: rule based

- ❖ Eliza (Weizenbaum, 1966)
 - ❖ Cobot in lambda moo (Isbell et al., 2000)
 - ❖ A.L.I.C.E (Wallace, 2008)- using AIML
 - ❖ Jabberwocky (Carpenter)
 - ❖ Pandorabots (Mitsuku)

```
=====
EEEEEEEEE L IIIIIIII ZZZZZZZZ AAA
E L I Z A A
E L I Z A A
EEEEEEEEE L IIIIIIII ZZZZZZZZ AAAA
E L I Z A A
EEEEEEEEE LLLLLLLL IIIIIIII ZZZZZZZZ A A
=====

ELIZA > (PLEASE TYPE IN ALL CAPS) WHAT'S YOUR NAME DEAR ?
AMIT
ELIZA > HI, I'M ELIZA, WHAT DO YOU WANT TO TALK ABOUT ?
AMIT > CAN I TALK ABOUT MY PROBLEM ?
ELIZA > SURE... !
AMIT > DUE TO ON GOING EXAMS IN COLLEGE, I HAVE A LOT OF STRESS.
ELIZA > PLEASE RELAX A LITTLE. SLEEP WELL.
AMIT > THANKS FOR YOUR ADVICE.
ELIZA > NO MENTION.
AMIT > BYE
ELIZA > BYE AND KEEP IN TOUCH...
=====
```



Types of systems: data-driven

1. **Response generation systems** that attempt to produce, word-by-word, a likely response given a conversational context using
 - e.g. (Deep) Neural Processing Models (Vinyals and Le, 2015, Shang et al., 2015, Li et al., 2016, Serban et al., 2016).
2. **Retrieval-based systems** that attempt to select a response from a (possibly large) list of utterances in a corpus
 - e.g. Next Utterance Classification (Lowe et al., 2016)
3. **The statistical machine translation (SMT) based method**
 - e.g. (Ritter et al., 2011).

Informative

Performance

Coherence

Interaction quality

Humous

Enjoyment

Fun

Conversational Depth

Chatty

Diversity

Acceptability

Usability

Appropriateness

Information Flow

Extrinsic vs Intrinsic

- ❖ **Extrinsic evaluations** assess the effect of a system on something that is external to it, in the real world
 - ❖ Task Success i.e. fulfils a user goal (e.g. book flights, find a restaurant)
- ❖ **Intrinsic evaluations** assess properties of systems in their own right
 - ❖ Automatic, e.g. BLEU
 - ❖ Hand-labelled, e.g. appropriateness, fluency

Redefining ‘Extrinsic’

1. User Task Success

- i.e. system helps the *user perform* a task through interaction (e.g., find restaurant information, find information on hotels, find treasure by giving instructions (GIVE))

2. Situated Task Success

- i.e. system fulfils a user’s situated goal by *performing an action* e.g. through 3rd party service provider (e.g. books flights, books restaurants, orders flowers)

3. User Internal Goal Fulfilment*

- i.e. system helps user to fulfil their *internal goal through interaction* (e.g. discuss politics, learn French, cheer themselves up by chatting)

4. User External Goal Fulfilment*

- i.e. system helps the user to fulfil their *external goal outside of the interaction* (e.g. cook a meal in the evening, finish their essay)

Ex. User and Situated Task Success

I'm looking for a child-friendly restaurant

OK, what kind of cuisine?

How about American?

Joe's diner is child-friendly and serves American food.

Can you book a table for 4 for lunch today?

Yes, sure no problem..hold on...that's it booked for you on TopTable. You can login to see the reservation

User goal: find a child-friendly restaurant



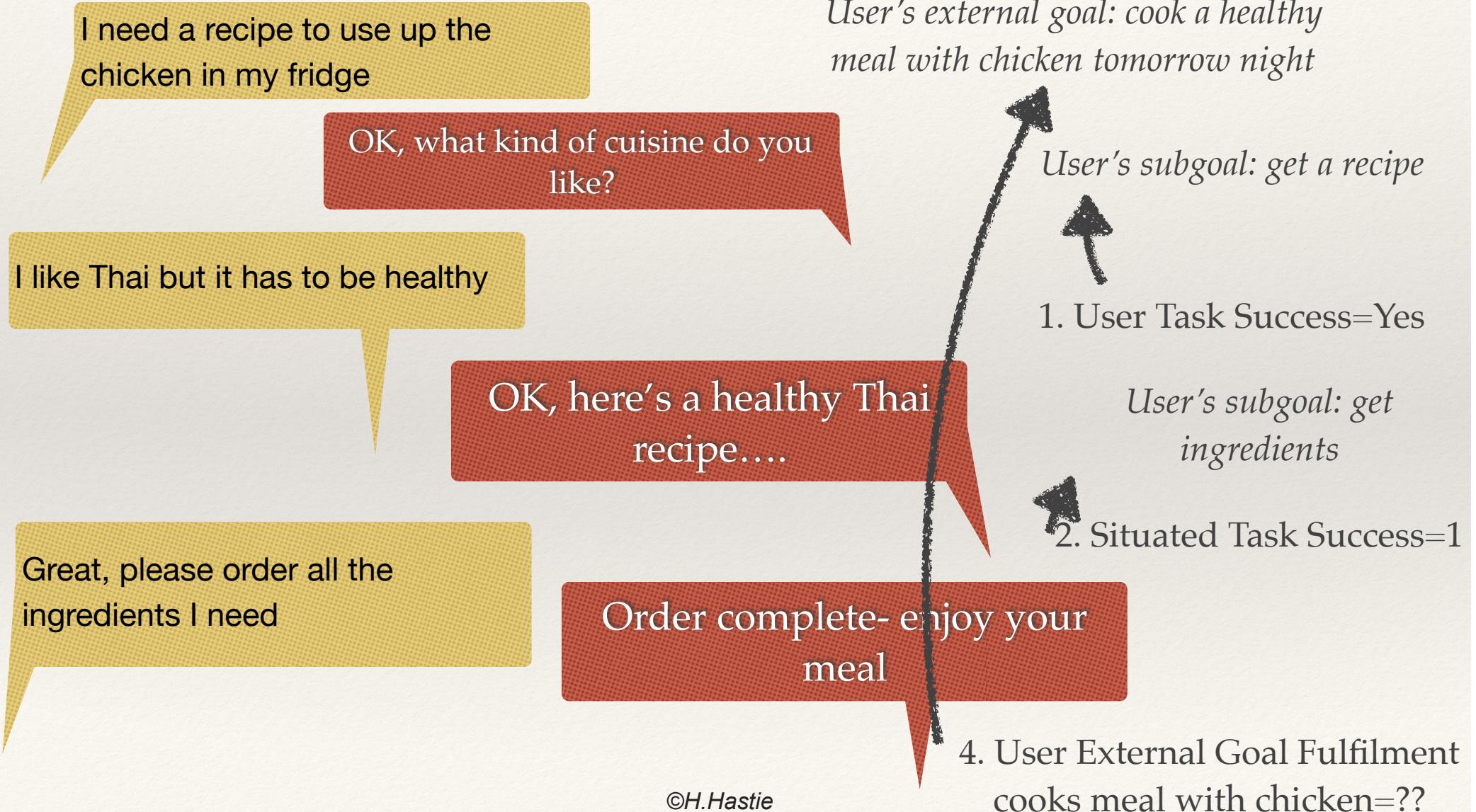
1. User Task Success=Yes

User goal: book a child-friendly restaurant

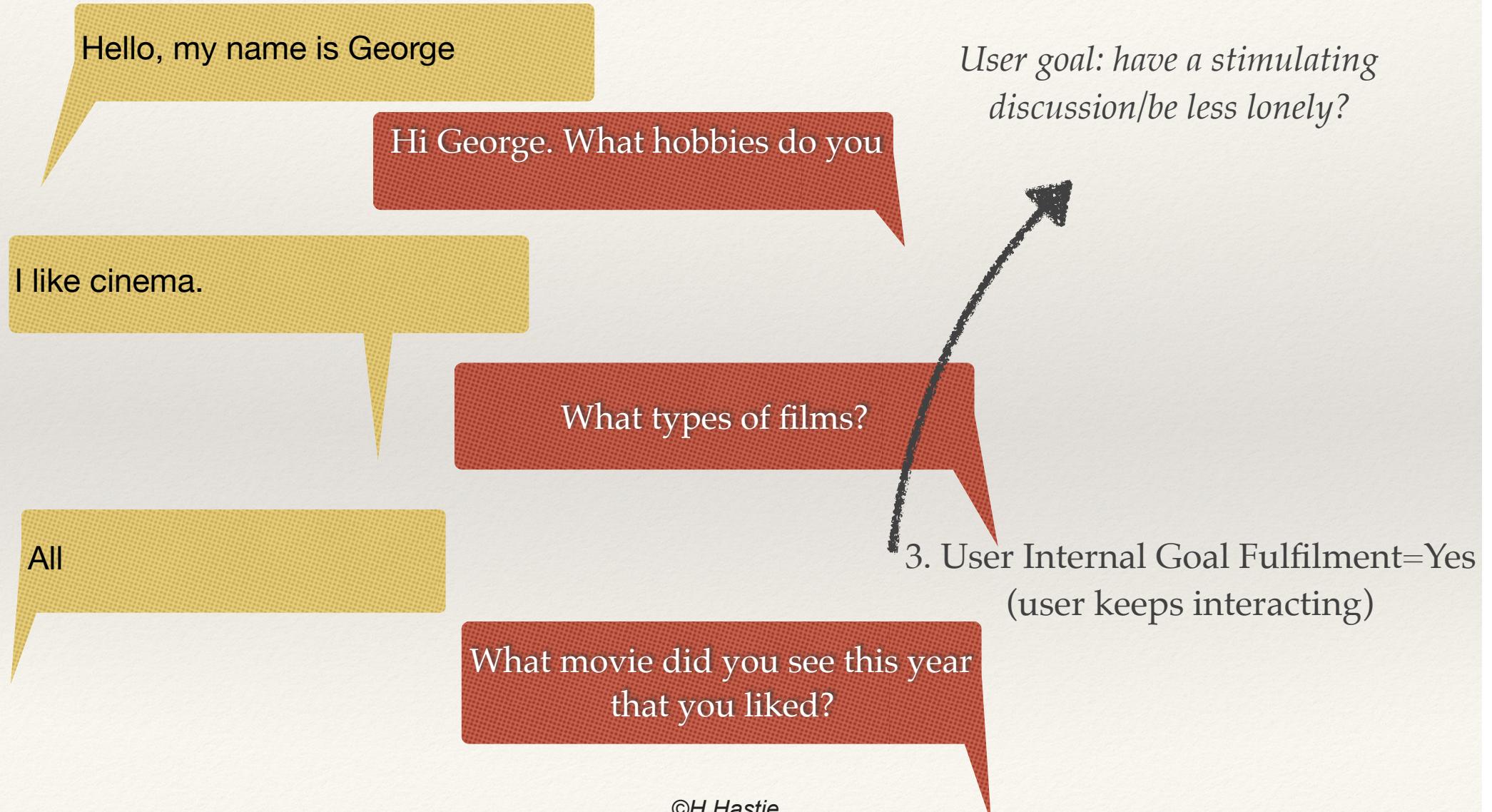


2. Situated Task Success=Yes

Ex. User External Goal Fulfilment



Ex. User Internal Goal Fulfilment



Redefining ‘Extrinsic’

1. User Task Success

- i.e. system helps the *user perform* a task through interaction (e.g., find restaurant information, find information on hotels, find treasure by giving instructions (GIVE))

2. Situated Task Success

- i.e. system fulfils a user’s situated goal by *performing an action* e.g. through 3rd party service provider (e.g. books flights, books restaurants, orders flowers)

3. User Internal Goal Fulfilment*

- i.e. system helps user to fulfil their *internal goal through interaction* (e.g. discuss politics, learn French, cheer themselves up by chatting)

4. User External Goal Fulfilment*

- i.e. system helps the user to fulfil their *external goal outside of the interaction* (e.g. cook a meal in the evening, finish their essay)

How measure?

How measure?

Evaluation Frameworks



INTERNATIONAL TELECOMMUNICATION UNION

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.851

(11/2003)

SERIES P: TELEPHONE TRANSMISSION QUALITY,

TELEPHONE INSTALLATIONS, LOCAL LINE

NETWORKS

Methods for objective and subjective assessment of
quality

Response var:
User
Satisfaction

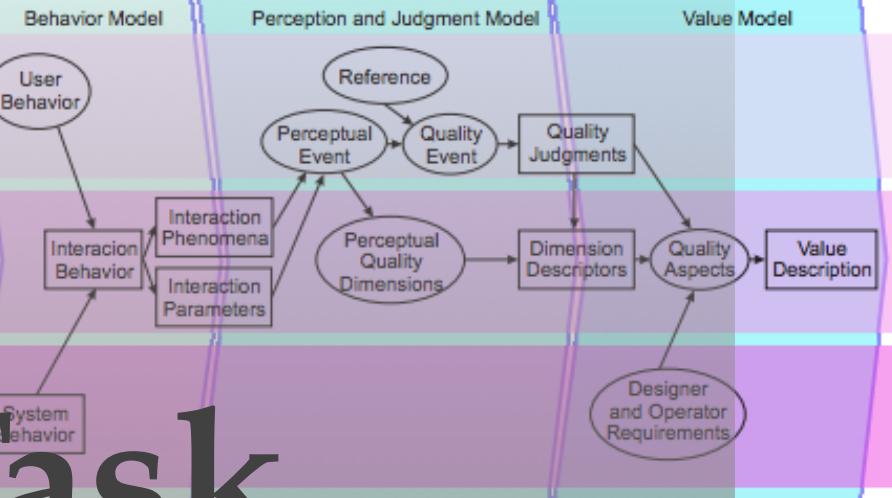
User Task

Maximise
Task Success

Efficiency
Measures
e.g. ASR
accuracy

Minimize
Costs

Qualitative
Measures, e.g.
repair ratio,



Möller and Ward (2008)

PARADISE (Walker et al. 1997)

©H.Hastie

Metrics by Approach

Approach	Human	Automatic	Estimated
Generative Approaches	Pairwise comparison (Li et al. 2016a, Vinyals & Le, 2015); next utterance rating; 5 turn 3rd party rating (Li et al., 2016b)	Diversity (Li et al. 2016a); BLEU, DeltaBLEU; Length metrics; Perplexity (Vinyals & Le, 2015); ROUGE (Xu et al., 2016); METEOR (Sordoni et al., 2015)	Appropriateness; Conversational depth; Information gain (Yu et al., 2016a)
Retrieval (NUC)	Pairwise comparison (Lowe et al., 2016)	Recall (Kaldec et al., 2015, Lowe et al., 2016); Precision (Al-Rfou et al., 2016)	
SMT	Pairwise comparison (Ritter et al., 2011)	BLEU (Ritter et al., 2011)	
WoZ	User Engagement, User engagement change, appropriateness (Yu et al., 2016b)	Count of mitigation strategies used (Yu et al., 2016b) •	

Evaluation Hierarchy

Mental
Model

Continued use, Trust,
Personality

Variety of topics,
triviality, responsiveness

System
Level

Engagement
Flow

Coherence, Information Flow, Conversational
Depth, Diversity, Consistency

Dialogue
Level

Response
Appropriateness

Response Appropriateness, Fluency, Coherence

Turn
Level

Subjective

Objective

Evaluation Hierarchy

Response
Appropriateness

Response Appropriateness, Coherence

Turn
Level

- Subjective
- Objective

TURN LEVEL SUBJECTIVE Response Relevance / Appropriateness

Example

Participant: How old are you?

TickTock: Apple.

Participant: How old are you?

TickTock: That's too big a question for me to answer.

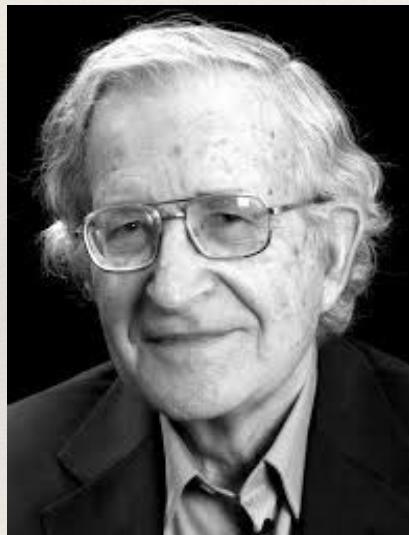
Participant: How is the weather today?

TickTock: Very good.

TURN LEVEL SUBJECTIVE Response Relevance / Appropriateness

- ❖ **Appropriateness/coherence with user utterance**
 - ❖ Ratings (Webb, 2010, Sordoni et al., 2015, Shang et al., 2015, Yu et al., 2016b)
 - ❖ Pairwise Comparison (Ritter et al., 2011, Li et al. 2016a, Vinyals & Le, 2015, Lowe et al., 2016)
- ❖ **Other Intrinsic measures**
 - ❖ Naturalness (Shang et al., 2015),
 - ❖ Fluency (Serban et al., 2016)
 - ❖ Adequacy/Clarity of reference (Li et al., 2016b)
 - ❖ Informativeness (Wen et al., 2015)

Expert vs Turkers vs Customers



TURN LEVEL OBJECTIVE

Response Appropriateness, Fluency, Coherence

- ❖ Challenges
 - ❖ Non-constrained dialogues,
 - ❖ No ‘one answer’: multiple ‘good’ responses (Sordoni et al., 2015),
 - ❖ Need to take into account the context of the turn / conversation.

TURN LEVEL OBJECTIVE

Response Appropriateness, Coherence

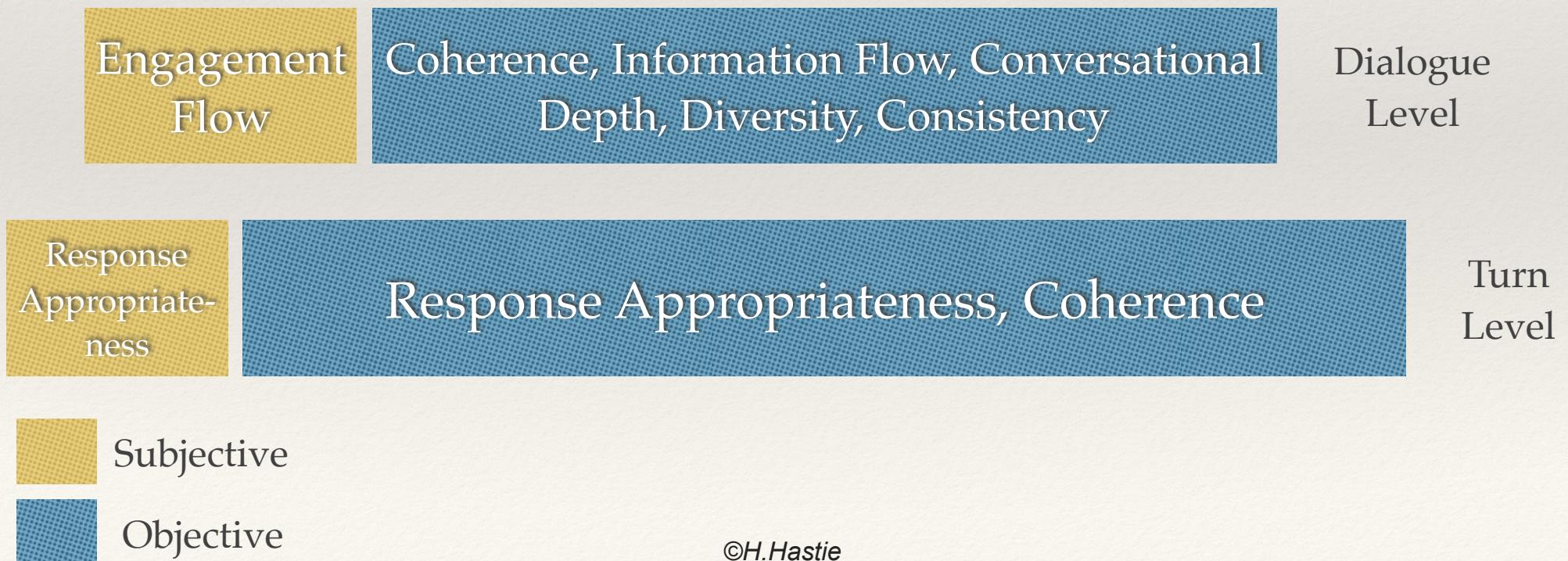
- ❖ Comparative metrics
 - ❖ BLEU, DeltaBLEU (Ritter et al., 2011, Liu et al., 2016); ROUGE (Xu et al., 2016); METEOR (Sordoni et al., 2015)
 - ❖ Perplexity (Vinyals & Le, 2015)
- ❖ Information Retrieval Metrics
 - ❖ Recall (Kaldec et al., 2015, Lowe et al., 2016)
 - ❖ Precision (Al-Rfou et al., 2016)

TURN LEVEL OBJECTIVE

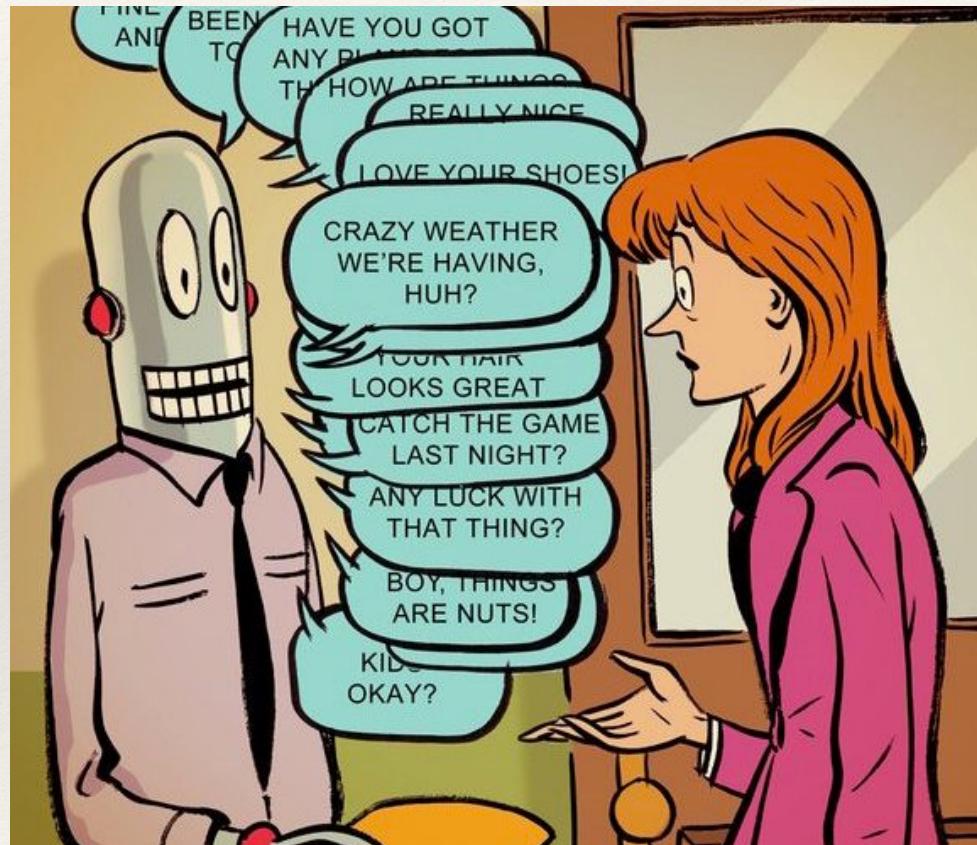
Response Appropriateness, Coherence

- ❖ Coherence: semantic / syntactic similarity metrics between user-system turn (Yu et al. 2016a)
- ❖ Confidence score (e.g. posterior probability for NN) (Yu et al., 2016a)
- ❖ Length metrics
- ❖ Approximating subjective ratings,
 - ❖ e.g. response appropriateness (73% accurate) via a classifier (Yu et al., 2016a)

Evaluation Hierarchy



DIALOGUE LEVEL SUBJECTIVE Engagement, Flow



“By engagement, we mean the process by which two (or more) participants *establish*, maintain and end their perceived *connection*. This process includes: initial contact, negotiating a collaboration, checking that other is still taking part in the interaction, evaluating whether to stay involved, and deciding when to end the connection.”

– Sidner, C. L., Kidd, C. D., Lee, C., & Lesh, N. (2003). *Where to Look: A Study of Human-Robot Engagement*. In ACM International Conference on Intelligent User Interfaces (IUI)

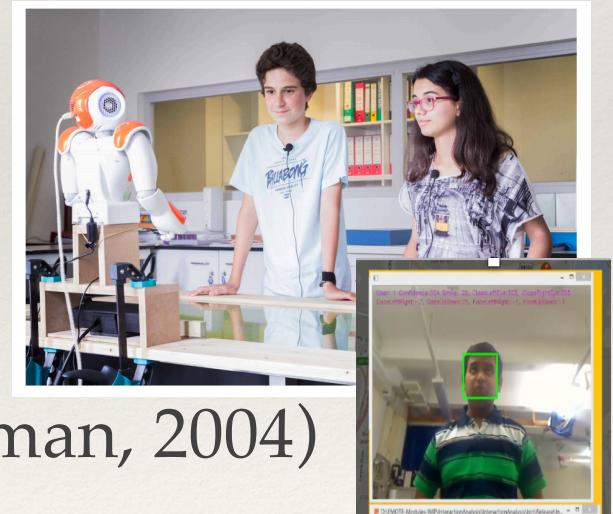
Engagement in Social Robotics

- ❖ Detect user affect including engagement and adapt interaction
- ❖ Robots have the advantage of :
 - ❖ having vision multiple modalities to detect engagement e.g. face, body, gesture, speech (Castellano et al., 2008), gaze (Sidney et al. 2003)
 - ❖ having language and gesture to mitigate disengagement



Engagement in Intelligent Tutoring Systems

- ❖ Engagement leads to motivation which leads to greater learning gain
- ❖ Detect affect and adapt pedagogical strategies / interaction
 - ❖ e.g. frustration, boredom (Castellano et al., 2013; Litman and Forbes, 2003)
- ❖ ITS have the advantage of:
 - ❖ having progress on a task as an indicator of engagement (Litman and Sillman, 2004)



Engagement as an Evaluation Measure

- ❖ Hand-annotated (1st person or third person) per turn or across whole dialogue (Yu et al., 2016b)
- ❖ Speech
 - ❖ Emotion recognition using speech using toolkits, e.g. OpenSSI, OpenEar, Multisense (Yu et al., 2015)
 - ❖ Fluency of interaction as indicators of engagement / attention e.g. user disfluencies, repairs, repetition (Shriberg, 1994)
- ❖ Language only: sentiment analysis

Engagement as reward signal

- ❖ Detect when we are successful in engaging the user and reinforce this behaviour using (Deep) Reinforcement Learning
 - ❖ Yu et al. (2015) maximise for Engagement as detected by MultiSense emotion recognition

Engagement and Flow

- ❖ Concept of Flow (named by Csíkszentmihályi)
 - ❖ is the mental state of operation in which a person performing an activity is fully immersed in a feeling of energized focus, full involvement, and enjoyment in the process of the activity.

Engagement and Flow cont.

- ❖ One must be involved in an activity with a clear set of *goals* and *progress*. This adds direction and structure to the task.
- ❖ The user should have clear and immediate feedback- allows them to adjust *their performance* to maintain the flow state.
- ❖ One must have *confidence* in one's ability to complete the task at hand. (Csíkszentmihályi et al., 2005)

Social system should support the user in his/her goal, give feedback/progress and increase user confidence

Engagement and Personalisation

- ❖ One way to make the user feel *involved* in the interaction and maintain flow is through personalisation
- ❖ Detect user style and adapt by matching linguistic style and lexical entrainment (Deutsch and Pechmann, 1982)



Personalisation in systems

- ❖ Rami Al-Rfou (2106) include the author ID in their response ranker to improve precision
- ❖ A Persona based Neural Model (Li et al., 2016c) performs entrainment and adaptation based on clusters of users (e.g. the system will respond to Brits differently to Americans).

DIALOGUE LEVEL SUBJECTIVE: Engagement, Flow

- ❖ Summary: How do we make systems more engaging?
 - ❖ Detect disengagement and adapt, e.g. through RL
 - ❖ Try to create Flow by making the user feel *involved* in the interaction
 - ❖ Personalisation may help keep users engaged

DIALOGUE LEVEL OBJECTIVE

Coherence, Information Flow, Conversational Depth, Diversity,

- ❖ **Information gain/flow**
 - ❖ e.g. number of unique words that are introduced into the conversation from both the system and the user. (Yu et al., 2016a)
 - ❖ *e.g. semantic dissimilarity between consecutive system turns (Li et al. 2016b)
- ❖ ***Semantic coherence** e.g. mutual information throughout dialogue (Li et al. 2016b)
- ❖ ***(System) Ease of answering**, e.g. negative log likelihood of responding to that utterance with a dull response (Li et al. 2016b)

DIALOGUE LEVEL OBJECTIVE

Coherence, Information Flow, Conversational Depth, Diversity,

- ❖ **Conversational depth** e.g. a classifier using as features the number of consecutive utterances that share the same topic ($\kappa = 0.45$) - 73% (Yu et al., 2016a)
- ❖ **Diversity of responses** e.g. number of distinct unigrams and bigrams (Li et al., 2016b)

Evaluation Hierarchy

Continued Use, Trust,
Personality

Variety of topics, triviality,
responsiveness

System
Level

Engagement
Flow

Coherence, Information Flow, Conversational
Depth, Diversity, Consistency

Dialogue
Level

Response
Appropriateness

Response Appropriateness, Coherence

Turn
Level

Subjective

Objective

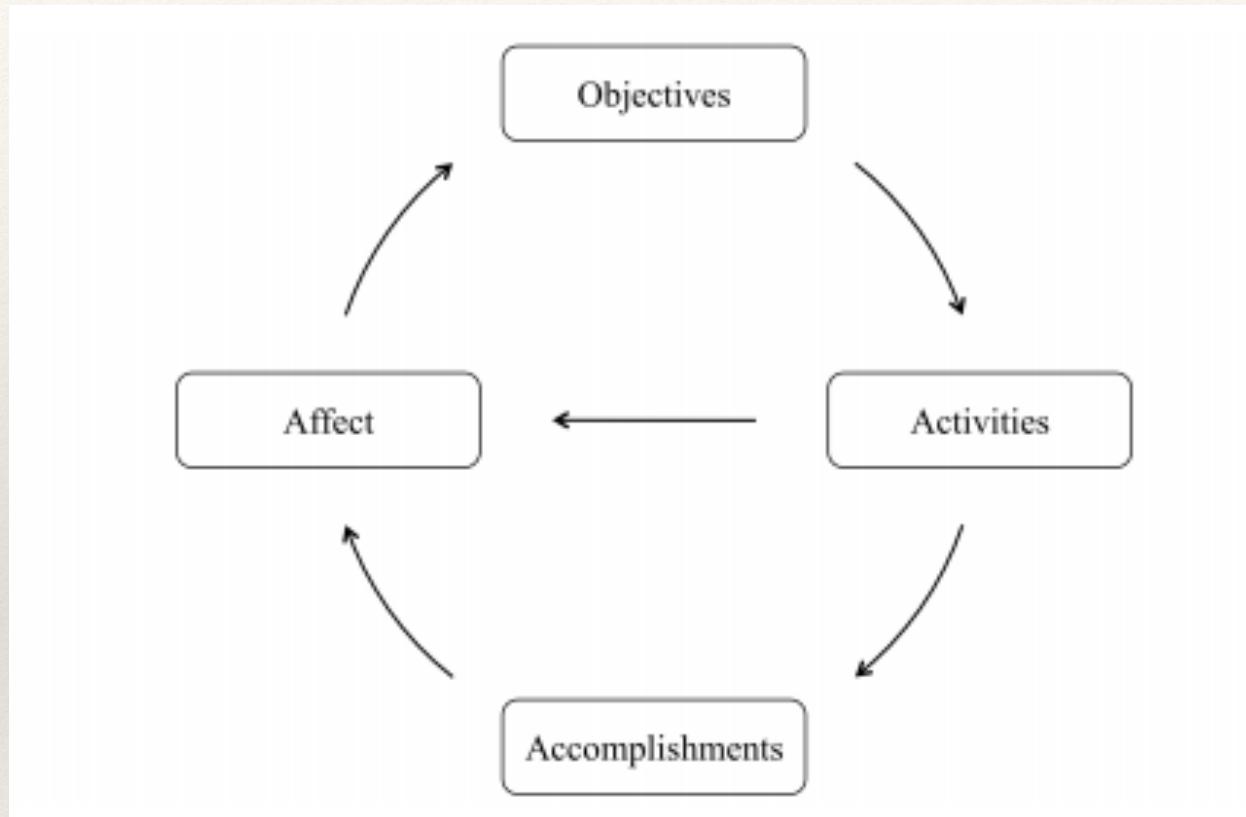
SYSTEM LEVEL SUBJECTIVE

Continuation Desire e.g. by repeated use

- ❖ Successful computer games have remarkable capability to:
 - ❖ draw people in (Jennett et al. 2008),
 - ❖ glue people to the game (Rigby and Ryan 2011), and
 - ❖ make people want to keep playing (Brown and Cairns 2004)

CONTINUATION DESIRE

SYSTEM LEVEL SUBJECTIVE Continuation Desire (Schoenau-Fog, 2011)



Key aspects are

- activities: interaction with game interface including absorbed in a story
- accomplishments: feeling of progression, completion

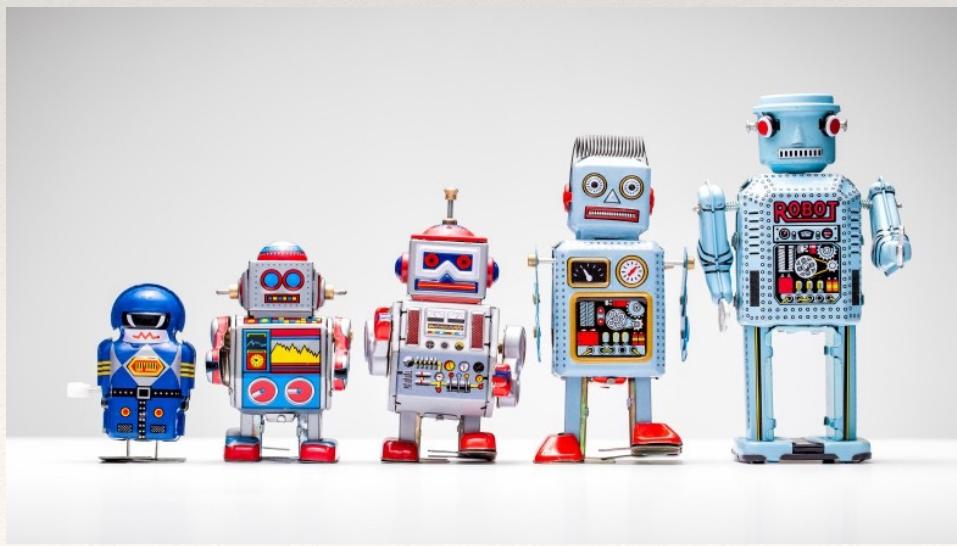
SYSTEM LEVEL SUBJECTIVE TRUST

- ❖ To continue using the system you have to trust it.
Consensus that *Trust* is multi-dimensional
 - ❖ e.g. seriousness, controllability, security, privacy and transparency (Liechtenstein et al. , 2009)
- ❖ Certain utterances make trigger increase / decrease in trust (Lumsden, 2009)
- ❖ Explanations can help increase transparency and also therefore trust (Lim et al., 2009)

Include explanations in social system to increase trust?

SYSTEM LEVEL SUBJECTIVE Personality

- ❖ User's perception of the system's personality can affect their subjective ratings



SYSTEM LEVEL SUBJECTIVE Personality

back

chatty [see definition of chatty](#) show

adj talkative

Relevance A-Z Complexity Length

Synonyms for chatty

Common Informal

adj talkative

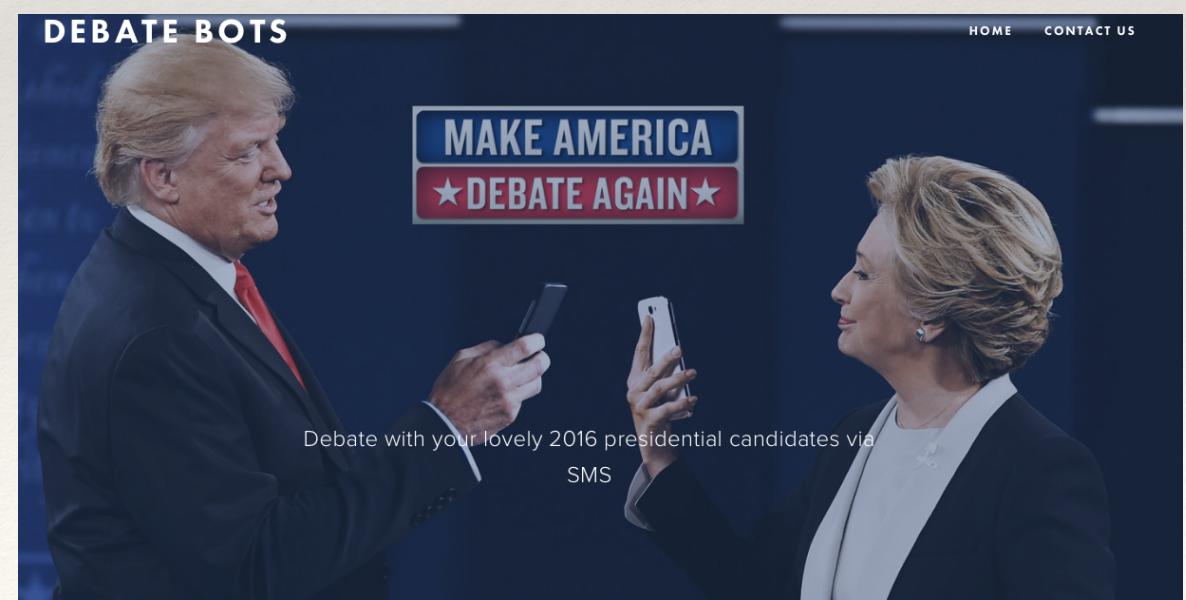
communicative	garrulous	loquacious	gabby	multielloquent
conversational	informal	colloquial	gossipy	spontaneous
friendly	intimate	familiar	loose-lipped	talky

What if Hillary and Donald were chatbots?

“On the one hand, you have an erratic, defective bot that keeps going off topic while throwing tantrums and repeating the same buzzwords over and over again. And he does it in a surprisingly hypnotizing fashion.

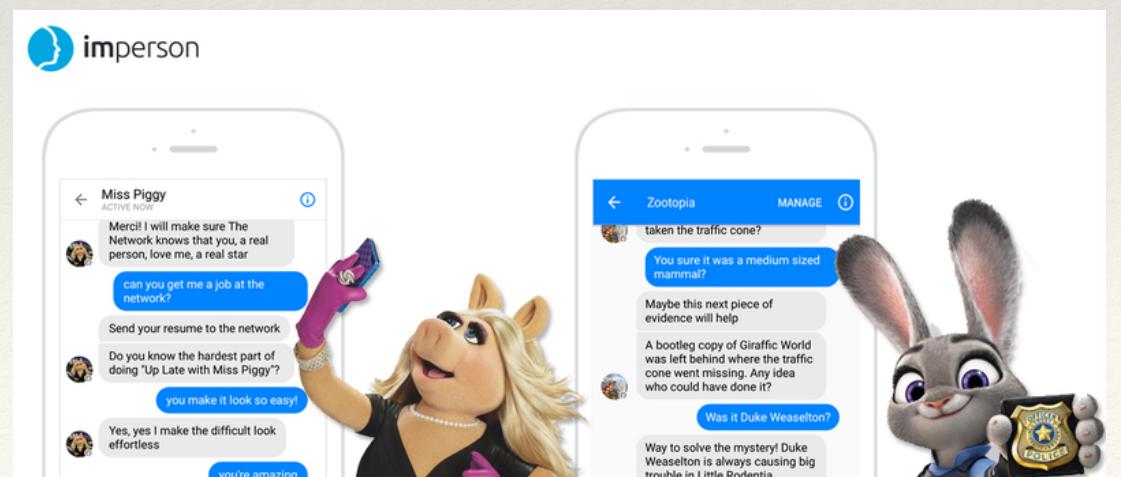
On the other hand, you have a perfectly crafted, overly trained bot that has pre-scripted fallback lines for just about every possible scenario that’s thrown at her (to further the engineering analogy, she’s been doing QA for over 30 years).”

Étienne Mérineau (Heyday)



SYSTEM LEVEL SUBJECTIVE Personality

- ❖ Rule-Based chatbots can include personality traits, e.g. for better customer service
- ❖ PERSONAGE Personality Natural Language Generation (Mairesse and Walker, 2007)
- ❖ For data-driven approaches: training data can influence the system personality
 - ❖ e.g. for trainable Surface realisers (Dethlefs et al., 2013)



SYSTEM OBJECTIVE

Ability to talk on a variety of topics, triviality, speed

- ❖ System should be able to:
 - ❖ Talk on a variety of topics and change smoothly between topics,
 - ❖ should be non-trivial to enhance flow and engagement topics? How define non-trivial?
 - ❖ be highly responsive.

Evaluation Hierarchy

Mental
Model

Continued use, Trust,
Personality

Variety of topics,
triviality, responsiveness

System
Level

Engagement
Flow

Coherence, Information Flow, Conversational
Depth, Diversity, Consistency

Dialogue
Level

Response
Appropriateness

Response Appropriateness, Fluency

Turn
Level

Subjective

Objective

We are all individuals.....



Man: ...I'm not.

Mental Models

- ❖ What the user believes about the system at hand
(Nielson)
- ❖ Mental Models in cognitive theory provide one view on how humans reason either
 - ❖ functionally (i.e. understanding what the system can do)
 - ❖ or structurally (i.e. understanding how it works).

Mental Models Cont.

- ❖ In HCI, designers make a user interface *communicate* the system's basic nature well enough that users form reasonably accurate (and thus useful) mental models

Mental Models Cont.

- ❖ What the user believes they know about a system strongly impacts
 - ❖ whether they use it,
 - ❖ whether they trust it,
 - ❖ how they use it.

Mental Models Cont.





Future Systems

- ❖ As systems learn new topics / skills on the fly and are constantly upgraded
- ❖ How can they subtly communicate a clear mental model to the user *on first use*
- ❖ Will they read minds and adjust interaction (i.e. Theory of Mind)

Challenges

- ❖ Loebner prize
- ❖ Alexa challenge
- ❖ WOCHAT
- ❖ SocialBot Challenge



Evaluation setting

- ❖ In the wild vs Amazon Turk
- ❖ Long term interaction
- ❖ Surviving in the real world and failing gracefully



Ethical Considerations

- ❖ The TayTweets chatbot that learned to say racist, bad things
- ❖ How can we validate sources of information we chat about?
- ❖ Socialbots are getting a bad rep: trying to get you in a social rapport and then sell you something or spread information
- ❖ Privacy of conversations
- ❖ How personal is too personal? Are chatbots contributing to the Echo chamber?

Summary

- ❖ To be a successful social chatbot, you have to be
 - ❖ engaging
 - ❖ respond appropriately and coherently but be diverse in your language
 - ❖ respond intelligently
 - ❖ draw people in and have them absorbed in the activity (flow)
 - ❖ portray a clear picture of what you can and can't do (mental model)
 - ❖ be able to discuss a wide range of topics
 - ❖ be ethical



Acknowledgements

- ❖ Dr. Verena Rieser and Amanda Cercas
- ❖ Colleagues at Interaction Lab including the Alexa team

References

- ❖ Al-Rfou, R., Pickett, M., Snaider, J., Sung, Y., Strope, B., & Kurzweil, R. (2016). Conversational Contextual Cues: The Case of Personalization and History for Response Ranking, 10. Retrieved from <http://arxiv.org/abs/1606.00372>
- ❖ Belz A. and H. Hastie (2014). Towards Comparative Evaluation and Shared Tasks for NLG in Interactive Systems. In Srinivas Bangalore and Amanda Stent (eds.) Natural Language Generation in Interactive Systems. Cambridge University Press
- ❖ Brown, E. and Cairns, P. (2004). "A grounded investigation of game immersion," in Extended Abstracts of the 2004 Conference on Human Factors in Computing Systems (Vienna, Austria, April 2004) ACM Press, pp. 1297-1300.
- ❖ Csikszentmihályi, M.; Abuhamdeh, S. & Nakamura, J. (2005), "Flow", in Elliot, A., Handbook of Competence and Motivation, New York: The Guilford Press, pp. 598–698
- ❖ Dethlefs, N. H.Hastie, H. Cuayáhuitl and O. Lemon (2013). Conditional Random Fields for Responsive Surface Realisation Using Global Features. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL). Sofia, Bulgaria
- ❖ K. Leichtenstern, E. Andr'e, and E. Kurdyukova (2010) "Managing user trust for self-adaptive ubiquitous computing systems," in MoMM'2010
- ❖ Ginevra Castellano, Ana Paiva, Arvid Kappas, Ruth Aylett, Helen Hastie, Wolmet Barendregt, Fernando Nabais and Susan Bull (2013) Towards Empathic Virtual and Robotic Tutors. In Proceedings of the 16th Conference on Artificial Intelligence in Education (AIED), Memphis, Tennessee, USA.
- ❖ J. Lumsden, "Triggering Trust: To What Extent Does the Question Influence the Answer When Evaluating the Perceived Importance of Trust Triggers?" in HCI, 2009.
- ❖ Jennett, C., A. L. Cox, P. Cairns, S. Dhoparee, A. Epps, T. Tijs, and A. Walton. "Measuring and defining the experience of immersion in games," in International Journal of Human-Computer Studies vol. 66, no. 9 (2008) Elsevier, pp. 641-61.
- ❖ Jeremić, Z., Jovanović, J., & Gašević, D. (2009). Evaluating an Intelligent Tutoring System for Design Patterns: the DEPTHS Experience. Educational Technology & Society, 12(2), 111–130.
- ❖ Kadlec, R., Schmid, M., & Kleindienst, J. (2015). Improved Deep Learning Baselines for Ubuntu Corpus Dialogs. Computation and Language. Retrieved from <http://arxiv.org/abs/1510.03753>
- ❖ Litman, D. J., & Silliman, S. (n.d.). ITSPoKE: An Intelligent Tutoring Spoken Dialogue System.
- ❖ Möller, S., & Ward, N. G. (n.d.). A Framework for Model-based Evaluation of Spoken Dialog Systems, 182–189.
- ❖ Rigby, S., & Ryan, R. Glued to games: How video games draw us in and hold us spellbound. ABC-CLIO, Santa Barbara, CA. 2011.
- ❖ Schoenau-Fog, H. (2011). The Player Engagement Process – An Exploration of Continuation Desire in Digital Games. In DiGRA 2011 Conference: Think Design Play.
- ❖ Sidner, C. L., Kidd, C. D., Lee, C., & Lesh, N. (2003). Where to Look: A Study of Human-Robot Engagement. In ACM International Conference on Intelligent User Interfaces (IUI)
- ❖ Yu, Z., Nicolich-Henkin, L., Black, A. W., & Rudnicky, A. I. (2016b). A Wizard-of-Oz Study on A Non-Task-Oriented Dialog Systems That Reacts to User Engagement. In SIGDIAL.
- ❖ Yu, Z., Xu, Z., Black, A. W., & Rudnicky, A. I. (2016a). Strategy and Policy Learning for Non-Task-Oriented Conversational Systems. In SIGDIAL .
- ❖ Charles Lee Isbell, Jr., Michael J. Kearns, Dave Ko- rmann, Satinder P. Singh, and Peter Stone. 2000. Cobot in lambdamoo: Asocial statistics agent. In Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on In- novative Applications of Artificial Intelligence, pages 36–41
- ❖ Dethlefs, N. et al. (2014) Cluster-based prediction of user ratings for stylistic surface realisation. In EACL.
- ❖ Information-Giving Agent. Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2015 - goDial), Aug 2015, Gothenburg, Sweden. Proceedings of the 19th Workshop on the Semantics and Pragmatics of Dialogue.
- ❖ Joseph Weizenbaum. 1966. Eliza: a computer program for the study of natural language communication between man and machine. Commun. ACM, 9:36–45, January.
- ❖ Li, J., Galley, M., Brockett, C., Gao, J., & Dolan, B. (2016a)). A Diversity-Promoting Objective Function for Neural Conversation Models, 110–119.
- ❖ Li, J., Monroe, W., Ritter, A., Galley, M., Gao, J., & Jurafsky, D. (2016b). Deep Reinforcement Learning for Dialogue Generation. In EMNLP.
- ❖ Li, J., Galley, M., Brockett, C., Gao, J., & Dolan, B. (2016c). A Persona-Based Neural Conversation Model. Acl, 10. Retrieved from <http://arxiv.org/abs/1603.06155>
- ❖ Liu, C.-W., et al. (2016). How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In EMNLP
- ❖ Lim, B. Y., Dey, A. K., & Avrahami, D. (2009). Why and Why Not Explanations Improve the Intelligibility of Context-Aware Intelligent Systems. In CHI (pp. 2119–2128).
- ❖ Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., & Pineau, J. (2016). On the Evaluation of Dialogue Systems with Next Utterance Classification. In SIGDIAL (pp. 264–269). Association for Computational Linguistics.
- ❖ Nadine Glas, Ken Pepin, Catherine Pelachaud. Engagement driven Topic Selection for an
- ❖ Ritter, A., Cherry, C., & Dolan, W. B. (2011). Data-driven response generation in social media. Proceedings of EMNLP
- ❖ Ritter, A., Cherry, C., & Dolan, W. B. (2011). Data-driven response generation in social media. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'11), 583–59
- ❖ Shang, Z. Lu, and H. Li. 2015. Neural responding machine for short-text conversation. arXiv preprint arXiv:1503.02364.
- ❖ Serban, V. Sordoni, A., Bengio, Y., Courville, A. and J. Pineau. 2016. Building end-to-end dia-logue systems using generative hierarchical neural network models. In Association for the Advance- ment of Artificial Intelligence (AAAI), 2016, pages 3776–3784.
- ❖ Webb, N., Benyon, D., Hansen, P. & Mival, O. (2010). Evaluating human-machine conversation for appropriateness. In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10).

Thank you for your attention