

CS-E4002 – Special Course in Computer Science:

Seminar on Computational Creativity

Lecture 4: Evaluating Computational Creativity

Anna Kantosalo, Christian
Guckelsberger & Tapio Takala



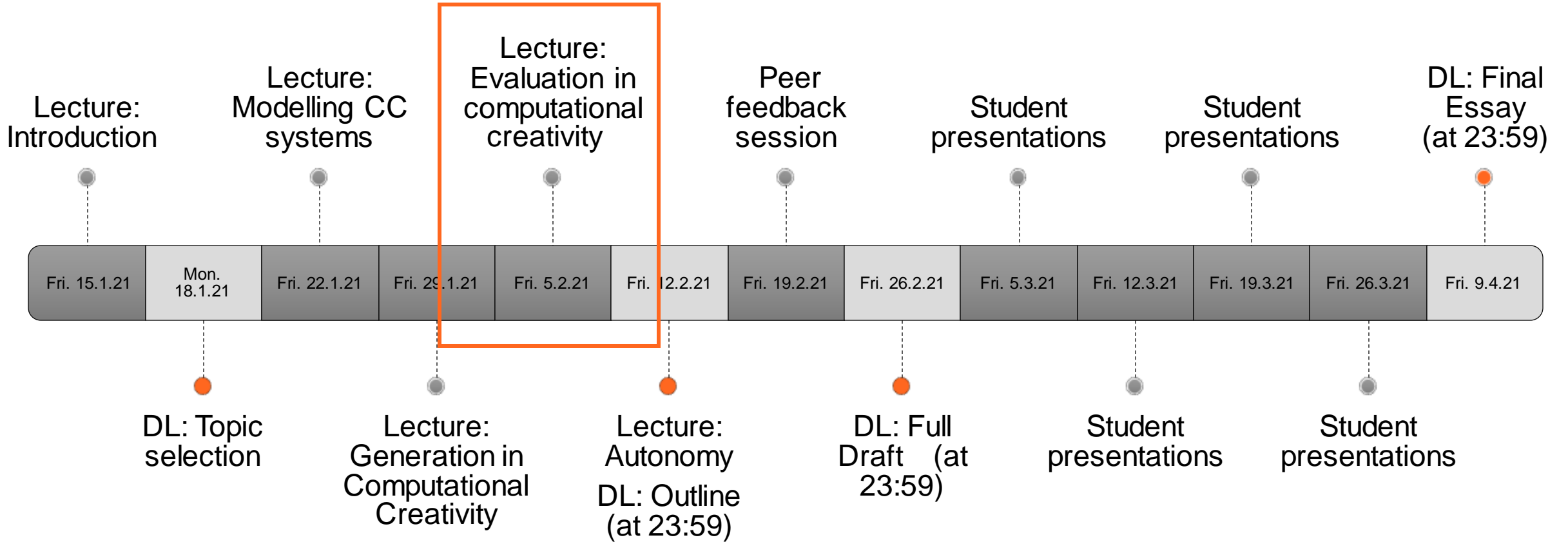
Aalto University
School of Science



Practicalities

Schedule

We are here



Lamp example

Inspired by "Spacetime constraints"
<https://doi.org/10.1145/378456.378507>



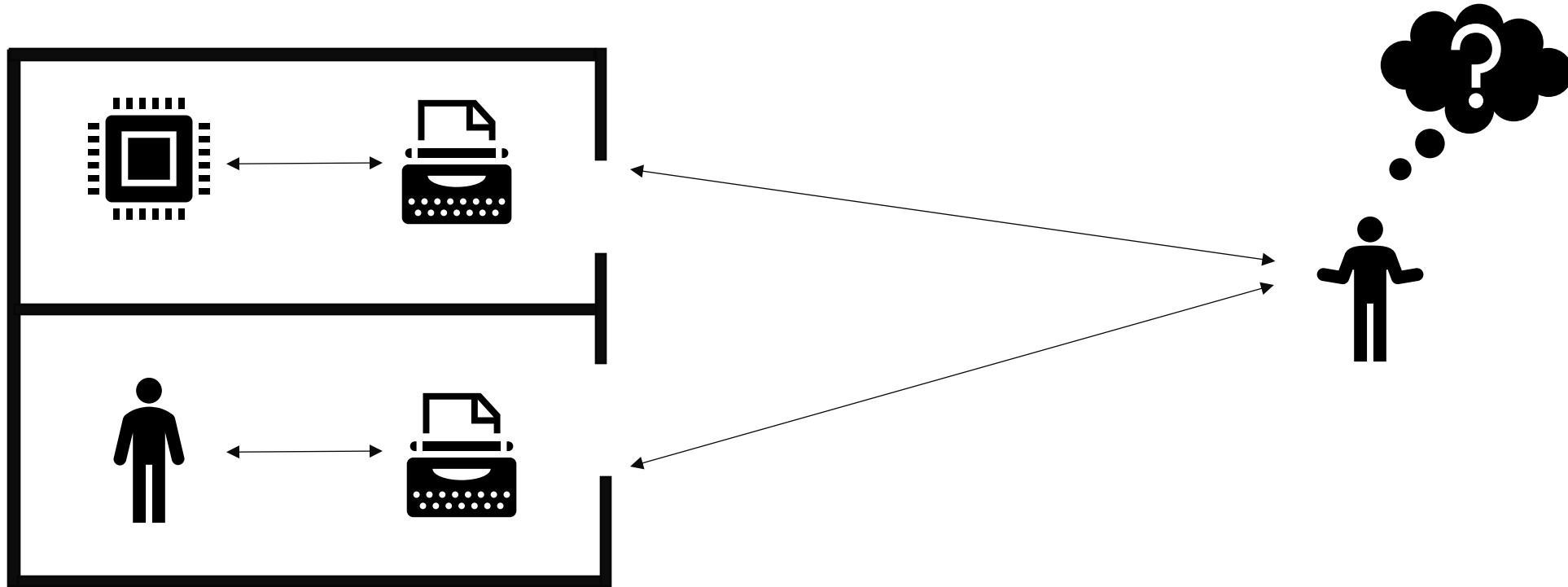
- Animated using motor programs developed by genetic programming
- Task for fitness test: move to the marked position
- Gradually adding constraints...

Evaluation

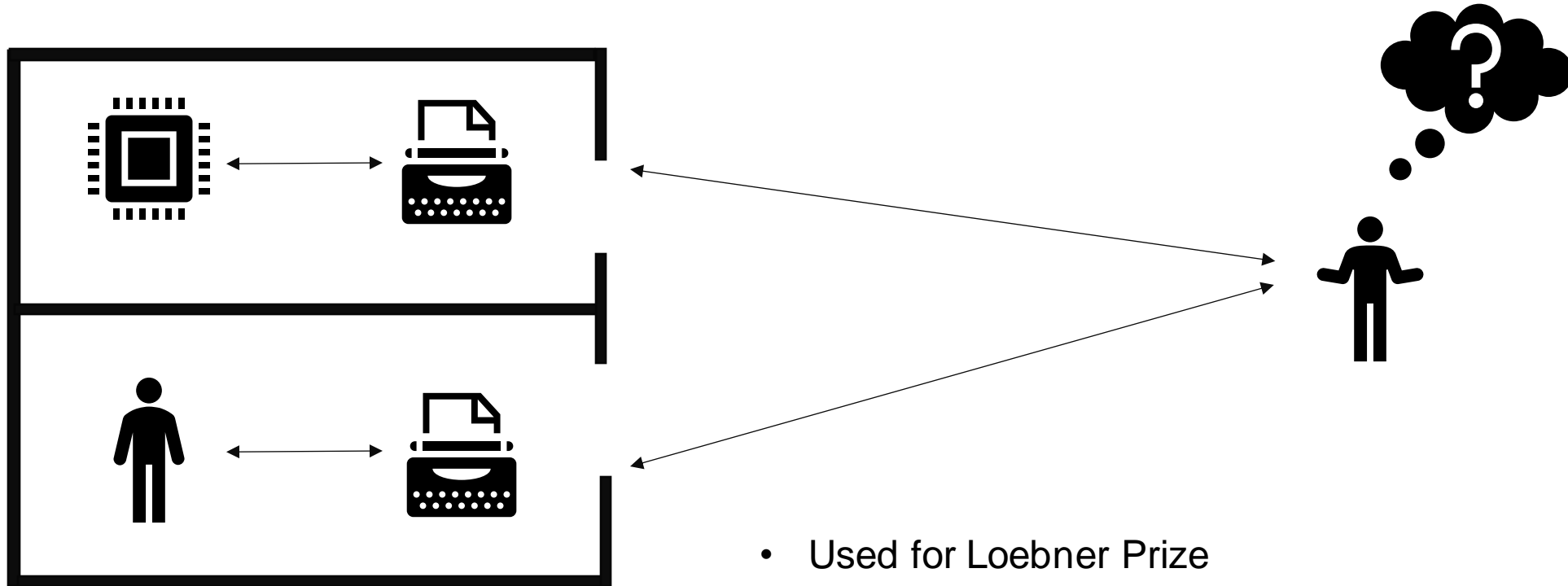
The Turing Test

"It is possible to do a little experiment ... even at the present stage of knowledge. It is not difficult to devise a paper machine which will play a not very bad game of chess. Now get three men as subjects for the experiment, A, B, and C. A and C are to be rather poor chess players. B is the operator who works the paper machine. (In order that he should be able to work it fairly fast it is advisable that he be both mathematician and chess player.) Two rooms are used with some arrangement for communicating moves, and a game is played between C and either A or the paper machine. C may find it quite difficult to tell which he is playing. (This is a rather idealized form of an experiment I have actually done.) (Turing 1948: 23)"

The Turing Test



The Turing Test



- Used for Loebner Prize
- A biased test: Humans tend to be labeled as computers (Copeland, 2000)

A Turing Test for Creativity?

A Turing Test for Creativity?

Boden, 2010:

“I will take it that for an ‘artistic’ program to pass the TT would be for it to produce artwork which was:

1. indistinguishable from one produced by a human being;

and/or

2. was seen as having as much aesthetic value as one produced by a human being.”

A Turing Test for Creativity?

Boden, 2010:

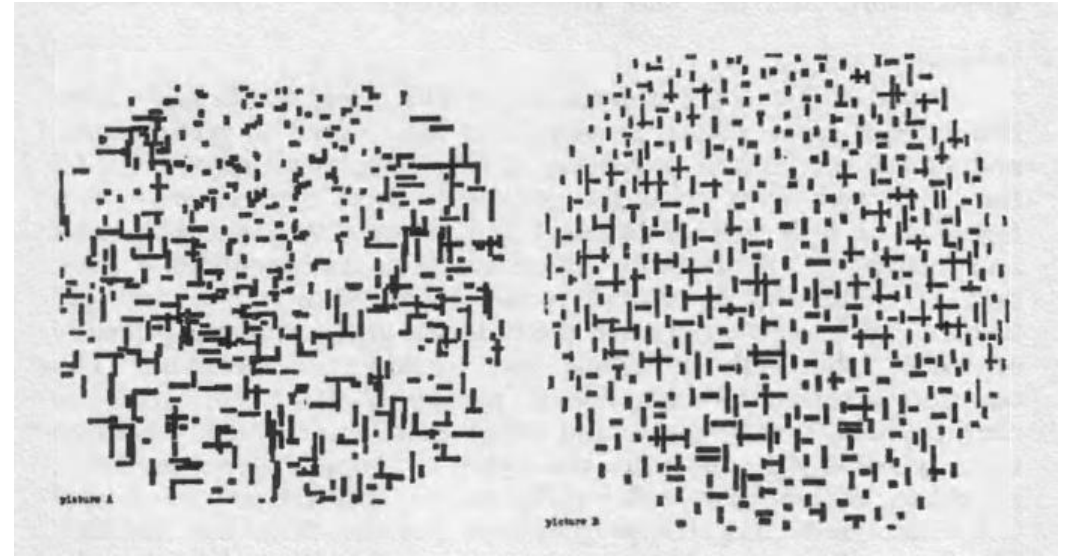
“I will take it that for an ‘artistic’ program to pass the TT would be for it to produce artwork which was:

1. indistinguishable from one produced by a human being;

and/or

2. was seen as having as much aesthetic value as one produced by a human being.”

Passed e.g. by a computer imitating Piet Mondrian's work in 1956 (Noll, 1966)



© Michael Noll & Rijkmuseum Kroller-Miiller

A Turing Test for Creativity?

Boden, 2010:

“I will take it that for an ‘artistic’ program to pass the TT would be for it to produce artwork which was:

1. indistinguishable from one produced by a human being;

and/or

2. was seen as having as much aesthetic value as one produced by a human being.”

Passed e.g. by AARON:



040502

AARON (2004)

pigment on paper

System designed by Harold Cohen

A Turing Test for Creativity?

Colton (2008) suggests that it is more relevant to ask **would someone buy computer generated art** than if they would be able to tell the difference between computer and human generated art.

A Turing Test for Creativity?

Colton, 2008 suggests that it is more relevant to ask **would someone buy computer generated art** than if they would be able to tell the difference between computer and human generated art.

Colton's 'test' has now been passed by the 'Portrait of Edmond de Belamy', a GAN generated portrait by the OBVIOUS art collective, sold at Christie's auction house for 432 000\$



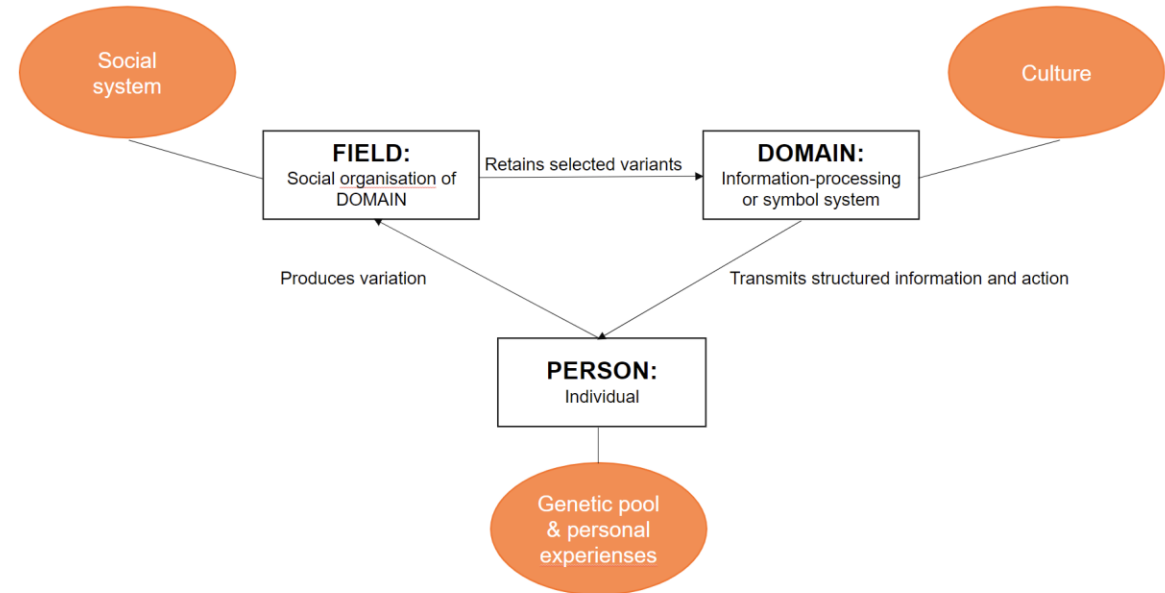
Is a Turing Test Useful?

Turing test for creativity Criticism

Pease & Colton, 2011

"[The Turing test] attempts to homogenise creativity into a single (human) style, does not take into account the importance of background and contextual information for a creative act."

The test is focused on the person perspective



Computational Creativity Evaluation

Agenda

- **WHY to evaluate?**
- **WHEN to evaluate?**
- **WHAT to evaluate?**

- **WHO should evaluate?**
- **HOW to evaluate?**

Analysing Your Topic

- **WHY to evaluate?** → Why was the system evaluated?
- **WHEN to evaluate?** → What is the maturity level of the project?
- **WHAT to evaluate?** → What has been evaluated and what criteria were used?
- **WHO should evaluate?** → Who did the evaluation?
- **HOW to evaluate?** → What kind of methods were used in the evaluation?

WHY to evaluate?

*”A comparative, scientific evaluation of creativity is **essential for progress in computational creativity**, not least to justify how creative a computational creativity system actually is.”*

- Jordanous, 2012

WHY to evaluate?

*”A comparative, scientific evaluation of creativity is **essential for progress in computational creativity**, not least to justify how creative a computational creativity system actually is.”*

- Jordanous, 2012

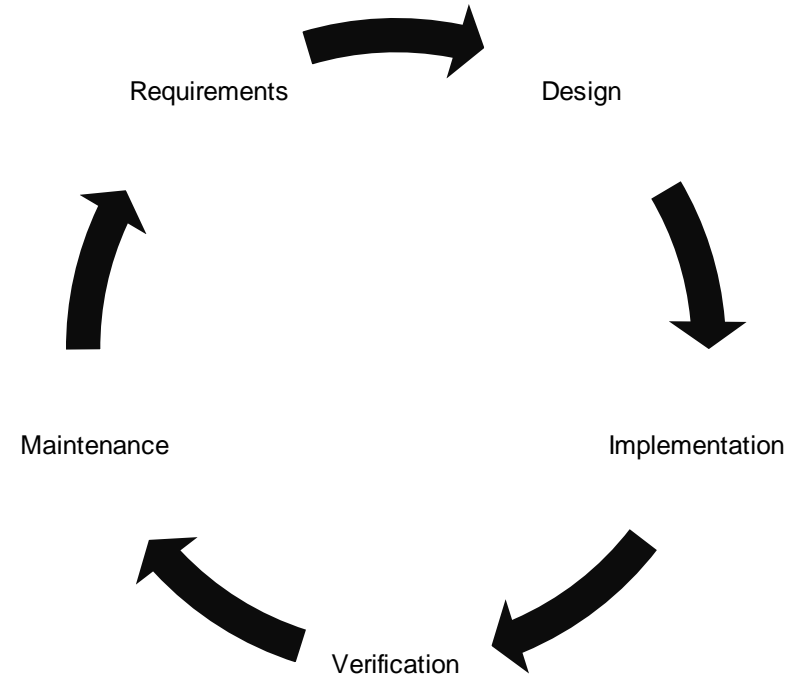
- **To measure improvement & show where improvement can be made**
- **To compare with other systems**
- **To argue for the creativity of a system**

WHEN to evaluate?

- **Evaluation is an iterative, on-going process**
- **Evaluation is a necessary part of scientific progress**
- **Systems should be evaluated in multiple stages during a project**

WHEN to evaluate?

- **Whatever software development model you use, you can identify projects in three different stages:**
 - **Early**
 - **On-Going**
 - **Finished**



WHEN to evaluate?

Questions for projects in early stages

- **What can be achieved with the chosen methodology?**
- **How is the planned system different from existing systems?**
- **What can be evaluated?**
- **What is considered a success?**

WHEN to evaluate?

Questions for on-going projects

- Does a system component do what it is intended to do?
- How can we boost its performance?

WHEN to evaluate?

Questions for finished projects

- **Does the system as a whole do what it was intended to do?**
- **How could we have boosted its performance further?**
- **How does the system compare to other similar systems?**

WHEN to evaluate?

Two types of evaluation questions

- **Summative evaluation**
 - Focus on a summary of the system's current level of creativity
- **Formative evaluation**
 - Constructive feedback on the strengths and weaknesses of a system to aid future development

WHAT to evaluate?

What aspects of the system should be evaluated?

Recall Rhodes' 4P's:

- **Person/Producer**
 - Qualities of the system producing creative artefacts
- **Process**
 - Algorithmic processes within, and interactions with the creative entity
- **Product**
 - The results of the creative process
- **Press/Environment**
 - The environment in which the creativity is situated

Adapted for CC by Jordanous (2016)

WHAT to evaluate?

What

ated?

Recall R

- **Person**

-

- **Process**

-

- **Product**

-

- **Press**

-

Product vs. Process vs. Producer:

- E.g. Ritchie (2007) advocates for viewing the product as the main evaluation target
- Colton (2008) and Colton, Charnley and Pease (2011) and others argue to include the process and producer in CC evaluation

entity

Adapted for CC by Jordanous (2016)

WHAT to evaluate?

Evaluation Criteria

- **Product**
 - Ritchie (2001, 2007):
 - Quality - Value of the product
 - Novelty - Has the product been invented before
 - Typicality - Is the product a typical example of similar products
 - Operationalised as computable ratings, such as the average typicality of produced items
 - Van der Velde et al. (2015):
 - Originality
 - Emotional value
 - Novelty/innovation
 - Intelligence
 - Skill

WHAT to evaluate?

Evaluation Criteria

- **Process & Product**
 - Colton, Charnley and Pease (2011): IDEA
 - Well-being rating - the personal hedonistic value of a creative act to an ideal audience

WHAT to evaluate?

Evaluation Criteria

- **Person/Producer**
 - Colton's (2008) Creative Tripod
 - Skillfulness - Ability to produce
 - Appreciativeness - Ability to evaluate the value of a product
 - Imaginativeness - Ability to produce novel items
- **Process**
 - Colton, Charnley and Pease (2011): FACE: Emphasizes a system's ability to report its process

WHAT to evaluate?

General Criteria by Jordanous (2012)

- Active involvement and persistence
- Dealing with uncertainty
- Domain competence
- General intellect
- Generation of results
- Independence and freedom
- Intention and emotional involvement
- Originality
- Progression and development
- Social interaction and communication
- Spontaneity/Subconscious processing
- Thinking and evaluation
- Value
- Variety, Divergence and Experimentation

WHAT to evaluate?

General Criteria by Jordanous 2012

- Active
- Dealin
- Doma
- Gene
- Gene
- Indep
- Intent

Evaluation criteria in CC have been operationalized to various degrees and only a few of them currently have standardized, statistically robust evaluation tools or objective measurements to accompany them.

opment
communication
ious
n
d

WHO evaluates?

Possible evaluators

- **Several entities can be involved in the evaluation process**
 - Internal evaluators (Evaluators involved in the project)
 - The system's designer
 - The creative system itself
 - External evaluators
 - Computational peers
 - Human audience (CC experts, domain experts, laymen)

WHO evaluates?

Combining evaluators

- **Different evaluation targets can be evaluated by different persons**
- **A combination of evaluators should be used to get more holistic, less biased and more useful results**

WHO evaluates?

Bias in evaluation

- People may have conscious or unconscious **bias for or against computers being creative** (Jordanous, 2012)
- People also have **different perceptions of creativity** and consider different things as creative (Jordanous, 2012)

WHO evaluates?

How to fight bias in evaluation

- Use a combination of evaluators and evaluation methods
- Use a large number of evaluators
- When using laymen as evaluators, pick them from varied backgrounds
- When comparing systems, choose comparisons that can be considered as representative of their domains
- When comparing computers to humans, do it blind
- Remember to randomize the order of evaluated systems, processes or artefacts
- **Keep the evaluation process transparent**, especially when doing a self-evaluation

How to evaluate?

- **Evaluation tools and methods can be divided into**
 - Quantitative vs. Qualitative
 - Summative vs. Formative
- **Many evaluation methods stem from or are supported by other disciplines: Psychology & Cognitive Science, Statistics, or Human-Computer Interaction ...**

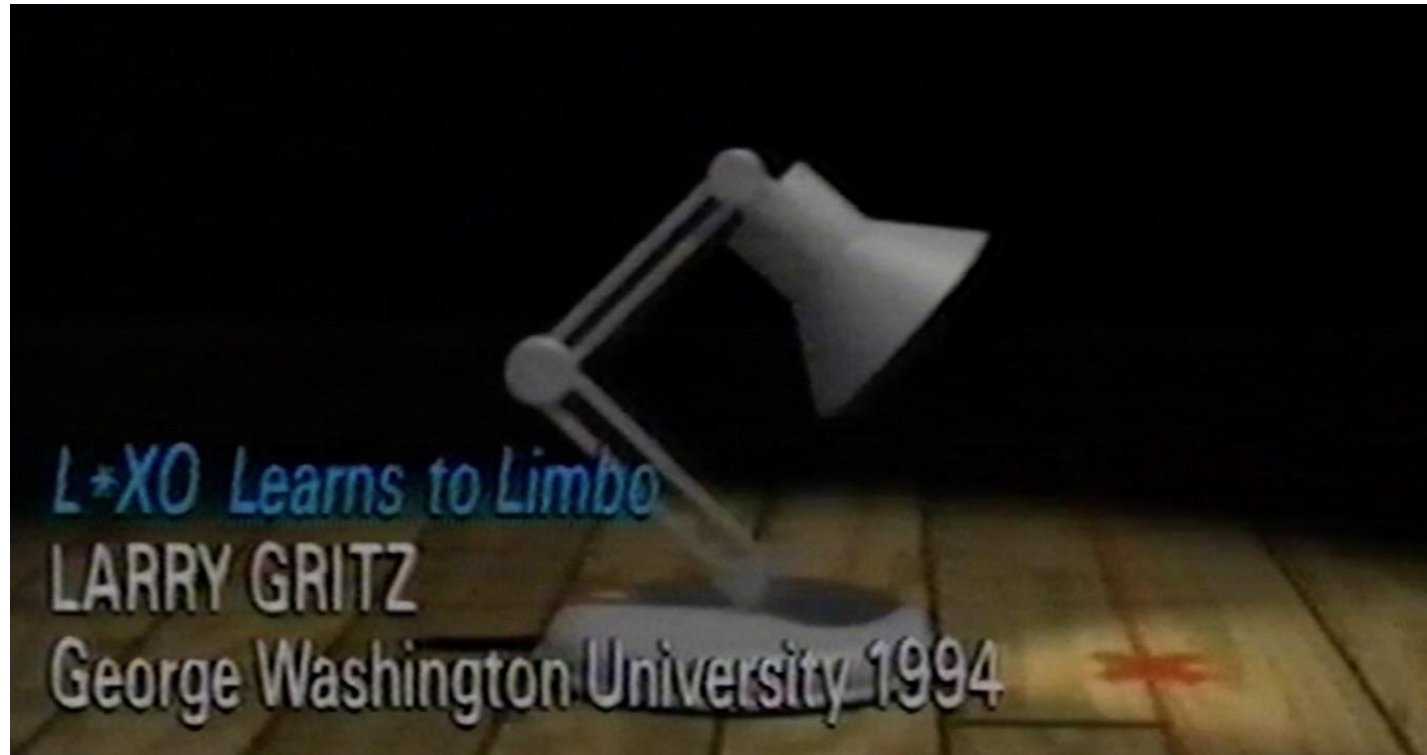
How to Evaluate

Colton, 2008: The creative tripod

- **Skill**
 - What kind of methods does the system use to produce art?
- **Appreciation**
 - Does the program have internal measures for appreciating what it produces?
- **Imagination**
 - Does the system produce novel outputs?

Practice example

– evaluate creativity of the lamp



https://miro.com/app/board/o9J_IWSax2A=/

How to evaluate?

SPECS: A standardised Procedure for Evaluating Creative systems (Jordanous, 2012)

- A domain independent way to **define an evaluation process** for a creative system

How to evaluate?

SPECS: A standardised Procedure for Evaluating Creative systems (Jordanous, 2012)

- **Step 1: Defining creativity**
 - Choose a definition the system should satisfy to be considered creative
 - What does it mean to be creative in general?
 - What aspects of creativity are important in the particular domain of the system?
 - What are you going to evaluate? Which P's are interesting to you?

How to evaluate?

SPECS: A standardised Procedure for Evaluating Creative systems (Jordanous, 2012)

- **Step 2: Identifying standards to test for**
 - Transform your definitions from step 1 to standards
 - Select criteria

How to evaluate?

SPECS: A standardised Procedure for Evaluating Creative systems (Jordanous, 2012)

- **Step 3: Test your creative system against the standards & report your results**
 - Operationalize the criteria, by devising tests to measure the performance of your system against the standard
 - Give more consideration to the standards representing aspects of creativity you have deemed more important for your domain
 - The test tools used depend on what is measured and what are the preferences, capabilities, equipment and facilities available to the researchers involved

How to Evaluate

Colton, Pease & Charnley, 2011:

FACE - evaluating process

E^g : an expression of a concept

E^p : a method for generating expressions of a concept

C^g : a concept

C^p : a method for generating concepts

A^g : an aesthetic measure

A^p : a method for generating aesthetic measures

F^g : an item of framing information

F^p : a method for generating framing information

How to Evaluate

Colton, Pease & Charnley, 2011:

FACE - evaluating process

E^g : an expression of a concept

E^p : a method for generating expressions of a concept

C^g : a concept

C^p : a method for generating concepts

A^g : an aesthetic measure

A^p : a method for generating aesthetic measures

F^g : an item of framing information

F^p : a method for generating framing information

Creative acts can be described as tuples e.g. $\langle E^g, C^g, A^g, F^g \rangle$.

Different acts can be compared:

$$\langle A^g \rangle < \langle A^g, C^g, E^g \rangle$$

(Read: A process evaluating, generating and expressing a concept is more creative than a system merely Evaluating it.)

Or different systems can be compared

- Average generated concept values
 - Best generated
 - Worst generated
 - Precision
- (portion of acceptably good concepts)
- Reliability (=best-worst)

Conclusions

- **Evaluation is critical to examine the creativity of computationally creative systems**
- **Evaluation is an essential requirement for good research**
- **To conduct a thorough evaluation, the researcher must identify**
 - when to evaluate
 - what to evaluate
 - who should evaluate
 - and how to conduct the evaluation
- **The evaluation must be thoroughly documented**

How to proceed with your essay?

- **Consider your example system or systems**
 - How would you define creativity in the domain of the system?
 - What aspects are required from the producer, the process, the product and the press?
 - What is the stage of the status of the system you are studying?
 - Is it early, on-going, or finished research?
 - What aspects of the system have the researchers evaluated?
 - Why?
 - When?
 - How?
 - And who did the evaluation?
 - What aspects can you evaluate yourself?
 - Use e.g. the FACE model, or Colton's Creative Tripod

How to read academic papers

How to read academic papers?

1. **Start with the abstract → Are you still interested?**
2. **Read the intro → Are you still interested?**
3. **Read the CONCLUSIONS → Is this still relevant?**
4. **Glance through the images & content → Is this still relevant?**
5. **Read the Background, Conclusions, Evaluation & Results**
6. **If it is a good paper, see the references for additional papers and/or try Google Scholar to find out who referenced this paper**

References

Boden, M. A. (2010). The Turing test and artistic creativity. *Kybernetes*.

Csikszentmihalyi, M. (1988). Society, culture, and person: A systems view of creativity. RJ Sternberg, ed. *The Nature of Creativity: Contemporary Psychological Perspectives*.

Colton, S. (2008, March). Creativity Versus the Perception of Creativity in Computational Systems. In *AAAI spring symposium: creative intelligent systems* (Vol. 8).

Colton, S., Charnley, J., Pease, A. (2011). Computational creativity theory: The FACE and IDEA descriptive models. In the Proceedings of the second international Conference on Computational Creativity, 2011. pp. 90-95

Copeland, B.J. *Minds and Machines* (2000) 10: 519.

Jordanous, A. (2016). Four PPPPerspectives on Computational Creativity in theory and in practice. *Connection Science*, 28(2), 194-216.

Jordanous, A. (2012). A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation*, 4(3), 246-279.

Liu, Y. T. (2000). Creativity or novelty?: Cognitive-computational versus social-cultural. *Design Studies*, 21(3), 261-276.

Noll, A. M. (1966). Human or machine: A subjective comparison of Piet Mondrian's "Composition with Lines"(1917) and a computer-generated picture. *The psychological record*, 16(1), 1-10.

Pease, A., & Colton, S. (2011, April). On impact and evaluation in computational creativity: A discussion of the Turing test and an alternative proposal. In *Proceedings of the AISB symposium on AI and Philosophy* (Vol. 39).

Ritchie, G. (2007). Some empirical criteria for attributing creativity to a computer program. *Minds and Machines*, 17, 67-99

Turing, A.M. (1948), 'Intelligent Machinery', National Physical Laboratory Report, in B. Meltzer and D. Michie, eds, *Machine Intelligence 5*, Edinburgh University Press (1969) as cited in Copeland, B.J. *Minds and Machines* (2000) 10: 519.

Aalto University School of Science
Val de Vede, M., Pease, A., Schmettow, M., & Nazareth, D. S. (2015). A semantic map for evaluating creativity. In Proceedings of the sixth International Conference on Computational Creativity. pp. 94-101

Ventura, D. (2017). How to Build a CC System. In *ICCC* (pp. 253-260).