MUSAE

Human-centred factory for a future technological sustainable development driven by arts

Human-machine interaction: best practices and tools

Francesco Semeraro, The University of Manchester, UK
Maria Bulgheroni, Ab.Acus, Italy

MANCHESTER 1824
The University of Manchester

Ab.Acus

MUSAE

European Commission

Training content

- Human-machine interaction experiment design (35 mins)
- Project workplan (10 mins)
- Q&A (10 mins)

User study design

In designing a human-robot interaction user study, you typically compare two or more different experimental conditions with each other, e.g. which settings of your system works best

You collect measurements regarding the interaction and then perform inferential statistical analysis to guarantee statistical significance of your results

But what are the key elements to consider in the design of the experiment?

Data management plan (refer to training from 06/11/2024)

---

https://www.amazon.co.uk/Statistical-inference-longer-Examples-exercises/dp/B0BF2KV8G2

F. Semeraro, J. Carberry, J. Leadbetter, and A. Cangelosi, “Good Things Come in Threes: The Impact of Robot Responsiveness on Workload and Trust in Multi-User Human-Robot Collaboration,” 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2024.

Experiment protocol

- A breakdown of an instance of your user study

- How many and who are your participants? Can they really give you meaningful answers?
  - If you can't find the ones you need, be ready to explain why your results can extend to the pool of interest

- Description of the experiment
  - How long does it last? One hour is already a lot
  - What will the participants have to do?
  - Online or in-person?

Experiment protocol: the task specifics

- One or multiple tasks?

- How many conditions?
  - You need to produce a feasible comparison term for your results
  - Try to go beyond two, but do something sensible to investigate

- Within subjects (participants get to experience all the conditions)
  - More data
  - Carryover effects

- Between subjects (participants experience only one condition, assigned randomly)
  - No carryover effects, but more demanding

More on user study design

- Risk assessment
  - Account for any source of hazard for the participants and how you plan to cover those
  - The participants are not in your head

- Participant information sheet and consent form
  - Broad description, not the whole detail
  - Explain how you are going to treat their data
  - Provide them with a point of contact

- Measuring instruments: 
  - Qualitative measurements
  - Quantitative measurements
  - Bias measurement

Qualitative measurements

- Typically performed by a psychologist
- 4 methodologies:
  - Concurrent Think Aloud: Ask questions during the experiment
  - Retrospective Think Aloud: Ask questions after the experiment
  - Concurrent Probing: Ask questions during the experiment only if certain things happen
  - Retrospective Probing: Ask questions after the experiment only if certain things happen

---

Priede, Camilla, and Stephen Farrall. "Comparing results from different styles of cognitive interviewing: 'verbal probing' vs 'thinking aloud'." *International Journal of Social Research Methodology* 14.4 (2011): 271-287.

Quantitative measurements

- The user performs a task with the system
- Define measurements specific to your task
- Effectiveness

$$TR = \frac{N_T}{T} 100 \text{ (completion rate)}$$

$$E = \frac{\sum_{j=1}^{R} \sum_{i=1}^{N} n_{ij}}{NR} \text{ (overall effectiveness)}$$

- Efficiency

$$P = \frac{\sum_{j=1}^{R} \sum_{i=1}^{N} \frac{n_{ij}}{t_{ij}}}{NR} \text{ (time-based efficiency)}$$

$$P_{te} = \frac{\sum_{j=1}^{R} \sum_{i=1}^{N} n_{ij} t_{ij}}{\sum_{j=1}^{R} \sum_{i=1}^{N} t_{ij}} 100 \text{ (overall relative efficiency)}$$

Usability: The SUS questionnaire

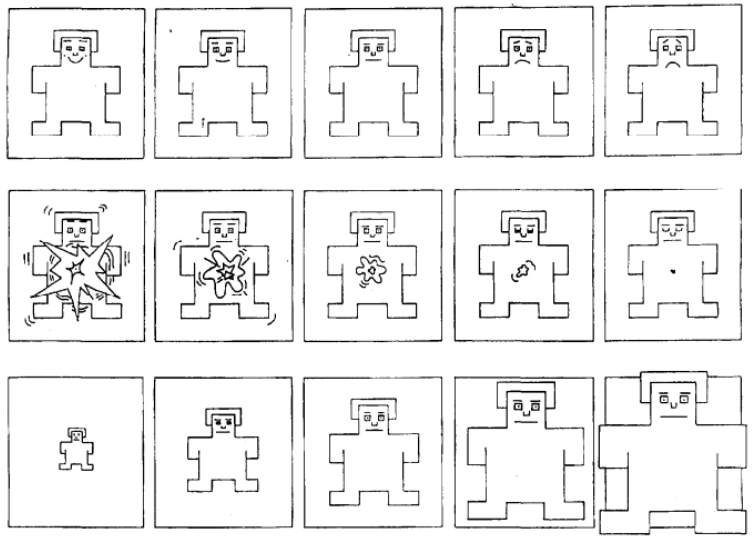
- It measures the perceived usability of a system
- It does not require to have another condition to compare against
- Threshold determines whether the system is usable

	The System Usability Scale Standard Version		Strongly Disagree				Strongly Agree
			1	2	3	4	5
1	I think that I would like to use this system frequently.		o	o	o	o	o
2	I found the system unnecessarily complex.		o	o	o	o	o
3	I thought the system was easy to use.		o	o	o	o	o
4	I think that I would need the support of a technical person to be able to use this system.		o	o	o	o	o
5	I found the various functions in this system were well integrated.		o	o	o	o	o
6	I thought there was too much inconsistency in this system.		o	o	o	o	o
7	I would imagine that most people would learn to use this system very quickly.		o	o	o	o	o
8	I found the system very awkward to use.		o	o	o	o	o
9	I felt very confident using the system.		o	o	o	o	o
10	I needed to learn a lot of things before I could get going with this system.		o	o	o	o	o

J. R. Lewis, ”The system usability scale: past, present, and future”, International Journal of Human–Computer Interaction, vol. 34, pp. 577-590, 2018.
A. Bangor, P. T. Kortum and J. T. Miller, ”An Empirical Evaluation of the System Usability Scale”, International Journal of Human–Computer Interaction, vol. 24, pp. 574–594, 2008
ISO. Ergonomic requirements for office work with visual display terminals (VDTs), Part 11, Guidance on usability (ISO 9241-11:1998E), 1998

Emotions: The SAM questionnaire

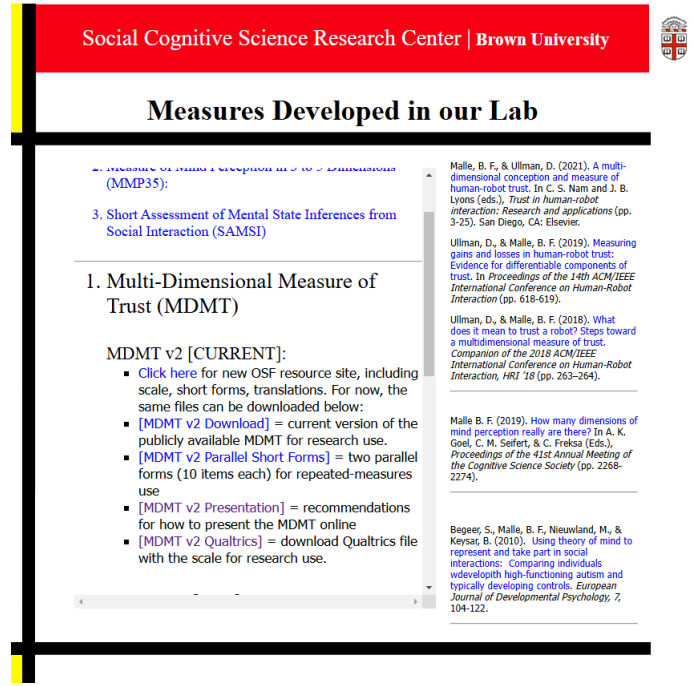
- The Self-Assessment Manikin probes the users regarding felt emotions
- Valence, arousal and dominance dimensions
- It is a picture-based questionnaire, so independent to the specific culture addressed



Bradley, Margaret M., and Peter J. Lang. "Measuring emotion: the self-assessment manikin and the semantic differential." *Journal of behavior therapy and experimental psychiatry* 25.1 (1994): 49-59.

Trust: The MDMT questionnaire

- The Multi-Dimensional Measure of Trust measures trust of users towards robots
- Also used for the human-human domain
- It measures two aspects of trust:
  - Performance trust: How do you trust the robot to be able to do its job?
  - Moral trust: How do you trust the moral sense of the robot?
- Suitable for repeated measurements



Mayer, Roger C., James H. Davis, and F. David Schoorman. "An integrative model of organizational trust." *Academy of management review* 20.3 (1995): 709-734.
Malle, Bertram F., and Daniel Ullman. "A multidimensional conception and measure of human-robot trust." *Trust in human-robot interaction*. Academic Press, 2021. 3-25.
https://research.clps.brown.edu/SocCogSci/Measures/

Trust: The TiA questionnaire

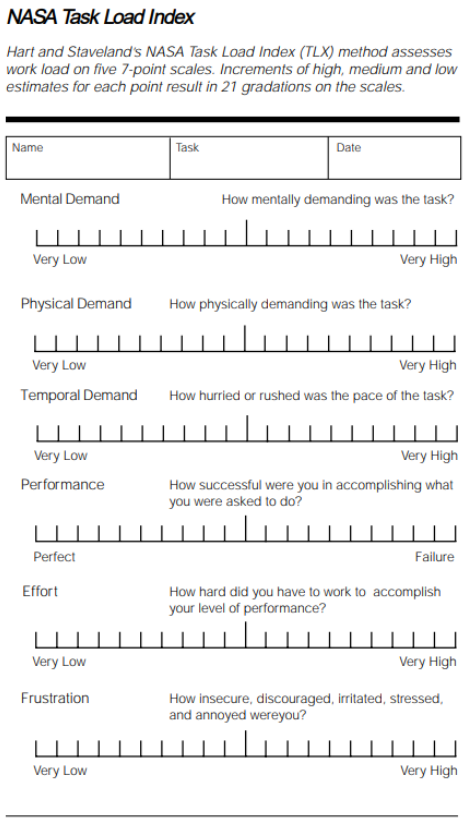
- Trust in Automation is a different questionnaire for measuring trust
- Only for performance trust, but in greater detail
- Not related necessarily to robotic agents, but automated agents

		Strongly disagree	Rather disagree	Neither disagree nor agree	Rather agree	Strongly agree	No response
1	The system is capable of interpreting situations correctly.	(1)	(2)	(3)	(4)	(5)	○
2	The system state was always clear to me.	(1)	(2)	(3)	(4)	(5)	○
3	I already know similar systems.	(1)	(2)	(3)	(4)	(5)	○
4	The developers are trustworthy.	(1)	(2)	(3)	(4)	(5)	○
5	One should be careful with unfamiliar automated systems.	(1)	(2)	(3)	(4)	(5)	○
6	The system works reliably.	(1)	(2)	(3)	(4)	(5)	○
7	The system reacts unpredictably.	(1)	(2)	(3)	(4)	(5)	○
8	The developers take my well-being seriously.	(1)	(2)	(3)	(4)	(5)	○
9	I trust the system.	(1)	(2)	(3)	(4)	(5)	○
10	A system malfunction is likely.	(1)	(2)	(3)	(4)	(5)	○
11	I was able to understand why things happened.	(1)	(2)	(3)	(4)	(5)	○
12	I rather trust a system than I mistrust it.	(1)	(2)	(3)	(4)	(5)	○
13	The system is capable of taking over complicated tasks.	(1)	(2)	(3)	(4)	(5)	○
14	I can rely on the system.	(1)	(2)	(3)	(4)	(5)	○
15	The system might make sporadic errors.	(1)	(2)	(3)	(4)	(5)	○
16	It is difficult to identify what the system will do next.	(1)	(2)	(3)	(4)	(5)	○
17	I have already used similar systems.	(1)	(2)	(3)	(4)	(5)	○
18	Automated systems generally work well.	(1)	(2)	(3)	(4)	(5)	○
19	I am confident about the system's capabilities.	(1)	(2)	(3)	(4)	(5)	○

Körber, M. “Theoretical considerations and development of a questionnaire to measure trust in automation.” In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018): Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics* (1st ed., pp. 13–30).

Workload: The NASA-TLX questionnaire

- The NASA Task Load Index measures the perceived workload of a task given to the user
- The scores can be weighted based on what the user think it is important



NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date

Mental Demand — How mentally demanding was the task?
Very Low — Very High

Physical Demand — How physically demanding was the task?
Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?
Very Low — Very High

Performance — How successful were you in accomplishing what you were asked to do?
Perfect — Failure

Effort — How hard did you have to work to accomplish your level of performance?
Very Low — Very High

Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?
Very Low — Very High

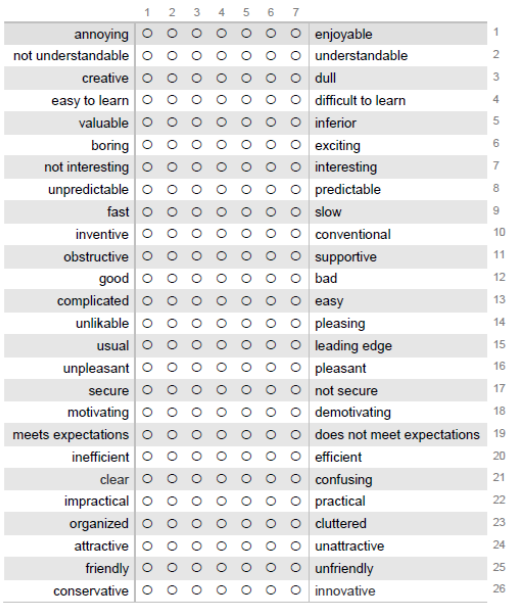
S. G. Hart, “Nasa-task load index (nasa-tlx); 20 years later”, Proceedings of the Human Factors and Ergonomics Society, pp. 904–908, 2006.
https://ntrs.nasa.gov/api/citations/20000021488/downloads/20000021488.pdf

User experience: The UEQ and AttrakDiff questionnaires

- Mainly meant to evaluate the users' perception on interacting with user interfaces
- They catch pragmatic and hedonistic qualities
- Pick the choice of words it suits you the most

Please provide your impressions of the product you have tested by check marking your impression on the scale between the terms offered in each line.

	1	2	3	4	5	6	7	
human	☐	☐	☐	☐	☐	☐	☐	technical
isolating	☐	☐	☐	☐	☐	☐	☐	connective
pleasant	☐	☐	☐	☐	☐	☐	☐	unpleasant
inventive	☐	☐	☐	☐	☐	☐	☐	conventional
simple	☐	☐	☐	☐	☐	☐	☐	complicated
professional	☐	☐	☐	☐	☐	☐	☐	unprofessional
ugly	☐	☐	☐	☐	☐	☐	☐	attractive
practical	☐	☐	☐	☐	☐	☐	☐	impractical
likeable	☐	☐	☐	☐	☐	☐	☐	disagreeable
cumbersome	☐	☐	☐	☐	☐	☐	☐	straightforward
stylish	☐	☐	☐	☐	☐	☐	☐	tacky
predictable	☐	☐	☐	☐	☐	☐	☐	unpredictable
cheap	☐	☐	☐	☐	☐	☐	☐	premium
alienating	☐	☐	☐	☐	☐	☐	☐	integrating
brings me closer to people	☐	☐	☐	☐	☐	☐	☐	separates me from people
unpresentable	☐	☐	☐	☐	☐	☐	☐	presentable
rejecting	☐	☐	☐	☐	☐	☐	☐	inviting
unimaginative	☐	☐	☐	☐	☐	☐	☐	creative
good	☐	☐	☐	☐	☐	☐	☐	bad
confusing	☐	☐	☐	☐	☐	☐	☐	clearly structured
repelling	☐	☐	☐	☐	☐	☐	☐	appealing
bold	☐	☐	☐	☐	☐	☐	☐	cautious
innovative	☐	☐	☐	☐	☐	☐	☐	conservative
dull	☐	☐	☐	☐	☐	☐	☐	captivating
undemanding	☐	☐	☐	☐	☐	☐	☐	challenging
motivating	☐	☐	☐	☐	☐	☐	☐	discouraging
novel	☐	☐	☐	☐	☐	☐	☐	ordinary
unruly	☐	☐	☐	☐	☐	☐	☐	manageable

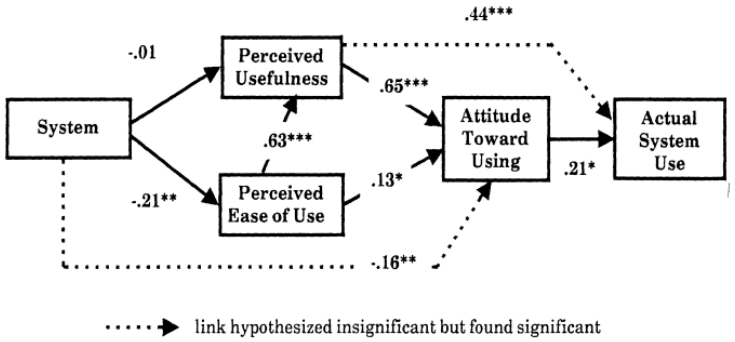


	1	2	3	4	5	6	7		
annoying	○	○	○	○	○	○	○	enjoyable	1
not understandable	○	○	○	○	○	○	○	understandable	2
creative	○	○	○	○	○	○	○	dull	3
easy to learn	○	○	○	○	○	○	○	difficult to learn	4
valuable	○	○	○	○	○	○	○	inferior	5
boring	○	○	○	○	○	○	○	exciting	6
not interesting	○	○	○	○	○	○	○	interesting	7
unpredictable	○	○	○	○	○	○	○	predictable	8
fast	○	○	○	○	○	○	○	slow	9
inventive	○	○	○	○	○	○	○	conventional	10
obstructive	○	○	○	○	○	○	○	supportive	11
good	○	○	○	○	○	○	○	bad	12
complicated	○	○	○	○	○	○	○	easy	13
unlikable	○	○	○	○	○	○	○	pleasing	14
usual	○	○	○	○	○	○	○	leading edge	15
unpleasant	○	○	○	○	○	○	○	pleasant	16
secure	○	○	○	○	○	○	○	not secure	17
motivating	○	○	○	○	○	○	○	demotivating	18
meets expectations	○	○	○	○	○	○	○	does not meet expectations	19
inefficient	○	○	○	○	○	○	○	efficient	20
clear	○	○	○	○	○	○	○	confusing	21
impractical	○	○	○	○	○	○	○	practical	22
organized	○	○	○	○	○	○	○	cluttered	23
attractive	○	○	○	○	○	○	○	unattractive	24
friendly	○	○	○	○	○	○	○	unfriendly	25
conservative	○	○	○	○	○	○	○	innovative	26

Hassenzahl M (2003) The Thing and I: Understanding the relationship between user and product. In: Lythe MA, Overbeeke K, Monk AF, Wright PC (eds) Funology. From Usability to Enjoyment, part of the Human-Computer Interaction Series. Kluwer Academic Publishers, Dordrecht, Netherlands, pp 31–42.
Schrepp, M.; Hinderks, A. & Thomaschewski, J. (2017). Construction of a benchmark for the User Experience Questionnaire (UEQ). International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 4, No. 4, pp. 40-44.

Your self-reported measurements (1/2)

- These are common aspects to investigate, but you can make your own questionnaire!
- Consider electronic surveys, e.g. through Qualtrics
- Models can help you out through the design
- Technology Acceptance (TAM) model
  - Perceived Usefulness
  - Perceived Ease of Use
  - User Acceptance



Davis, Fred D., R. P. Bagozzi and P.R. Warshaw. "Technology acceptance model." *J. Manag. Sci.* 35, 8 (1989)

Your self-reported measurements (2/2)

- TAM2 and TAM3 account for other influencing factors, like:
  - Social Influence
  - Facilitating Conditions

- UTAUT model
  - Performance Expectancy
  - Effort Expectancy

- Cronbach's Alpha

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma_{y_i}^2}{\sigma_y^2}\right)$$

Venkatesh, V., & Davis, F. D. "A theoretical extension of the technology acceptance model: Four longitudinal field studies." Management Science, 46(2), 186-204 (2000).
Venkatesh, Viswanath, and Hillol Bala. "Technology acceptance model 3 and a research agenda on interventions." *Decision sciences* 39.2 (2008): 273-315.
Khechine, Hager, Sawsen Lakhal, and Paterne Ndjambou. "A meta-analysis of the UTAUT model: Eleven years later." *Canadian Journal of Administrative Sciences/Revue Canadienne des Sciences de l'Administration* 33.2 (2016): 138-152.
Cronbach, Lee J. "Coefficient alpha and the internal structure of tests." *Psychometrika*. 16 (3). Springer Science and Business Media LLC: 297–334 (1951).

Bias measurement

- Measuring bias allows you to understand whether the result is conditioned upon the disposition of the participants to the structure of the experiment itself
- To administer before your experiment
- Compulsory if you are running a between-subjects study, you might not need it if you are running a within-subjects study
- Not a detriment to your study, but an opportunity to better understand what happened

M. Romeo, P. E. McKenna, D. A. Robb, et al., “Exploring theory of mind for human-robot collaboration,” RO-MAN 2022 - 31st IEEE International Conference on Robot and Human Interactive Communication, pp. 461–468, 2022.

The NARS and PTT questionnaires

- Negative Attitude Towards Robots measures mistrust in robots
  - Interaction
  - Influence
  - Emotions

- Propensity to Trust in Technology does not refer to robots necessarily

- Can be administered both to understand if the bias is related to robotics or more generally to technology

Item No.	Questionnaire Items
1	I would feel uneasy if robots really had emotions.
2	Something bad might happen if robots developed into living beings.
3	I would feel relaxed talking with robots. a)
4	I would feel uneasy if I was given a job where I had to use robots.
5	If robots had emotions, I would be able to make friends with them. a)
6	I feel comforted being with robots that have emotions. a)
7	The word "robot" means nothing to me.
8	I would feel nervous operating a robot in front of other people.
9	I would hate the idea that robots or artificial intelligences were making judgments about things.
10	I would feel very nervous just standing in front of a robot.
11	I feel that if I depend on robots too much, something bad might happen.
12	I would feel paranoid talking with a robot.
13	I am concerned that robots would be a bad influence on children.
14	I feel that in the future society will be dominated by robots.

1. Generally, I trust technology.
2. Technology helps me solve many problems.
3. I think it's a good idea to rely on technology for help.
4. I don't trust the information I get from technology. (R)
5. Technology is reliable.
6. I rely on technology.

Nomura, Tatsuya, et al. "Measurement of negative attitudes toward robots." *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems* 7.3 (2006): 437-454.
Schneider, T. R., Jessup, S. A., Stokes, C., Rivers, S., Lohani, M., McCoy, M.: The influence of trust propensity on behavioral trust. Poster session presented at the meeting of Association for Psychological Society, Boston, MA (2017, May).

Setting the Stage: Objectives and Scope

- **Project Objectives:**
  - What you're building and why: detail SMART (Specific, Measurable, Achievable, Relevant and Time bound) objectives
  - Key priorities: Usability, Sustainability, Ethics, Data privacy

- **Scope of Work:**
  - Technical Scope
  - Creative Scope
  - Collaboration Areas

Report your frame according to the work done in the concept feasibility phase

Breaking It Down: Workplan Structure

Split your work in workpackages (well defined portions of work). On a 6 months time scale 3-4 workpackages are fine. Each work package is described in terms of:

- Objectives
- Tasks (small chunks of work)
- Responsible person/team
- Required resources
- Expected output

Tips:

  - Synchronize creative and technical activities
  - Think of 15 days iterations (completion of small pieces of work)

Who Does What? (roles and responsibilities)

- **Technical Team:**
  - Developers (which activities)
  - Testers (which activities)
  - Data Specialists (which activities)

- **Artist role** (which activities)

- **Collaboration dynamics** (joint activities)

Timeline (Gantt chart)

- **Timeline:**
  - design and planning (M1)
  - iterative prototyping(M2-M5)
  - iterative testing (M3-M6)

THANK YOU

MUSAE

European Commission