Scientific challenges, practical methodologies and policy perspectives for trustworthy AI

Emilia Gómez (emilia.gomez-gutierrez@ec.europa.eu)

Work with the HUMAINT team

https://ai-watch.ec.europa.eu/humaint

European Commission

Outline

1. Intro
2. EU approach for trustworthy AI
3. Human behaviour and machine intelligence

About myself

https://emiliagomez.com/

1. Interdisciplinary background – engineering & music.
2. Information Retrieval/Recommender systems – audio & music.
3. Impact of algorithms on human behaviour.

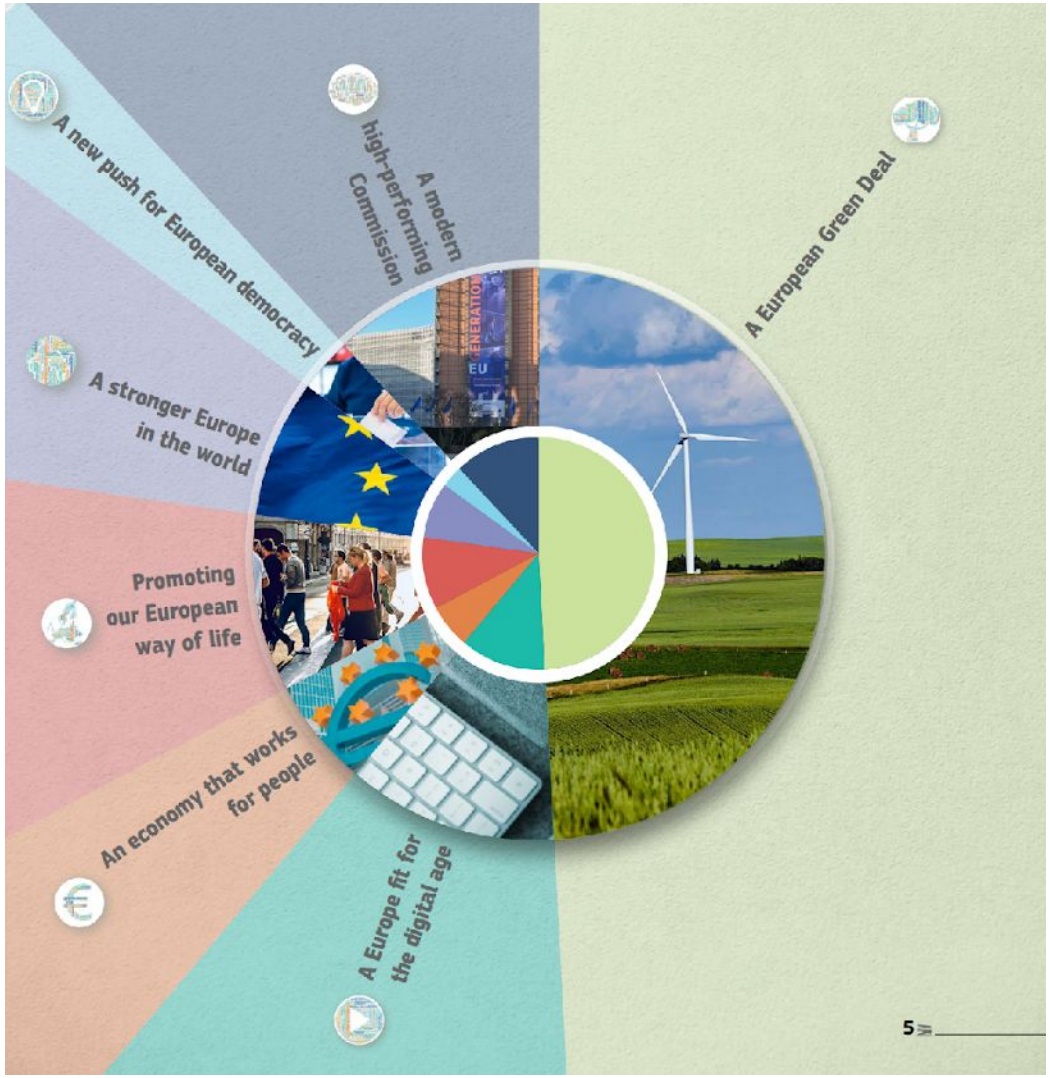
ISMIR

European Commission

Joint Research Centre

Role: provide evidence-based, scientific and technical support to European policies.

6 countries (**Brussels-Belgium**, Geel-Belgium, **Ispra-Italy**, Karlsruhe-Germany, Petten-Netherlands, **Seville-Spain**).



https://joint-research-centre.ec.europa.eu/index_en
@EU_ScienceHub

European Commission

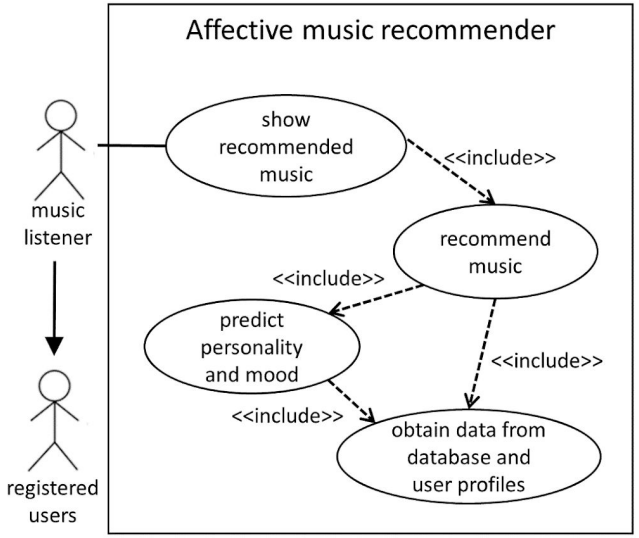
Ethical Guidelines + Legal Requirements + Standards + Engineering practices = Trustworthy AI

EU AI policies

European Centre for Algorithmic Transparency

https://ai-watch.ec.europa.eu/humaint

Socio-technical systems



USE CASE	**Show recommended music**	
Context of use	The user is subscribed to a music platform, which recommends the most appropriate and enjoyable tracks according to her personality and current mood. Personality and mood are predicted based on the data in the user's profile (voluntary provided by the user) and the historical music data she has listened to. The system also takes into account the music tracks played by other users with a similar profile to make recommendations. The user accesses the music platform through an application installed in her mobile phone.	
Intended purpose	Recommend a list of songs to the user according to her personality, current mood and music preferences.	
Application areas	Entertainment and leisure	
User	Music listener	
Target persons	**Person**	**Description**
	Registered users	Other users registered on the platform and whose profile and music preferences are used to make recommendations.
Success end condition	A list of 20 recommended music tracks is shown to the user in the application's graphic interface.	
Failure protection	A default personality- and mood-neutral list of 20 songs is shown to the user in the application's graphic interface.	
Trigger	The user presses the "recommend music" button in the application.	
Main course	**Step**	**Action**
	1	The application calls the recommender algorithm.
	2	The current mood of the user is predicted based on her profile information and recently played songs.
	3	The personality of the user is predicted based on her profile information and historical music playlists.
	4	The recommender ranks songs according to predicted mood, personality and music playlists of other registered users with similar profile.
	5	The application displays the 20 top-ranked recommended tracks for the user.
Extensions	**Step**	**Branching action**
	2a	If no song has been played yet, the system assigns the user a neutral mood.
	3a	If there is no historical music data, personality prediction is based on the user's profile information exclusively.
Misuses	The recommender shall not propose pieces of music pre-conceived to exploit vulnerabilities, manipulate, distort or induce certain emotions or behaviour in users, e.g. for marketing purposes.	

Hupont, I., & Gomez, E. (2022). Documenting use cases in the affective computing domain using Unified Modeling Language. Affective Computing and Intelligence Interaction https://arxiv.org/abs/2209.09666v1

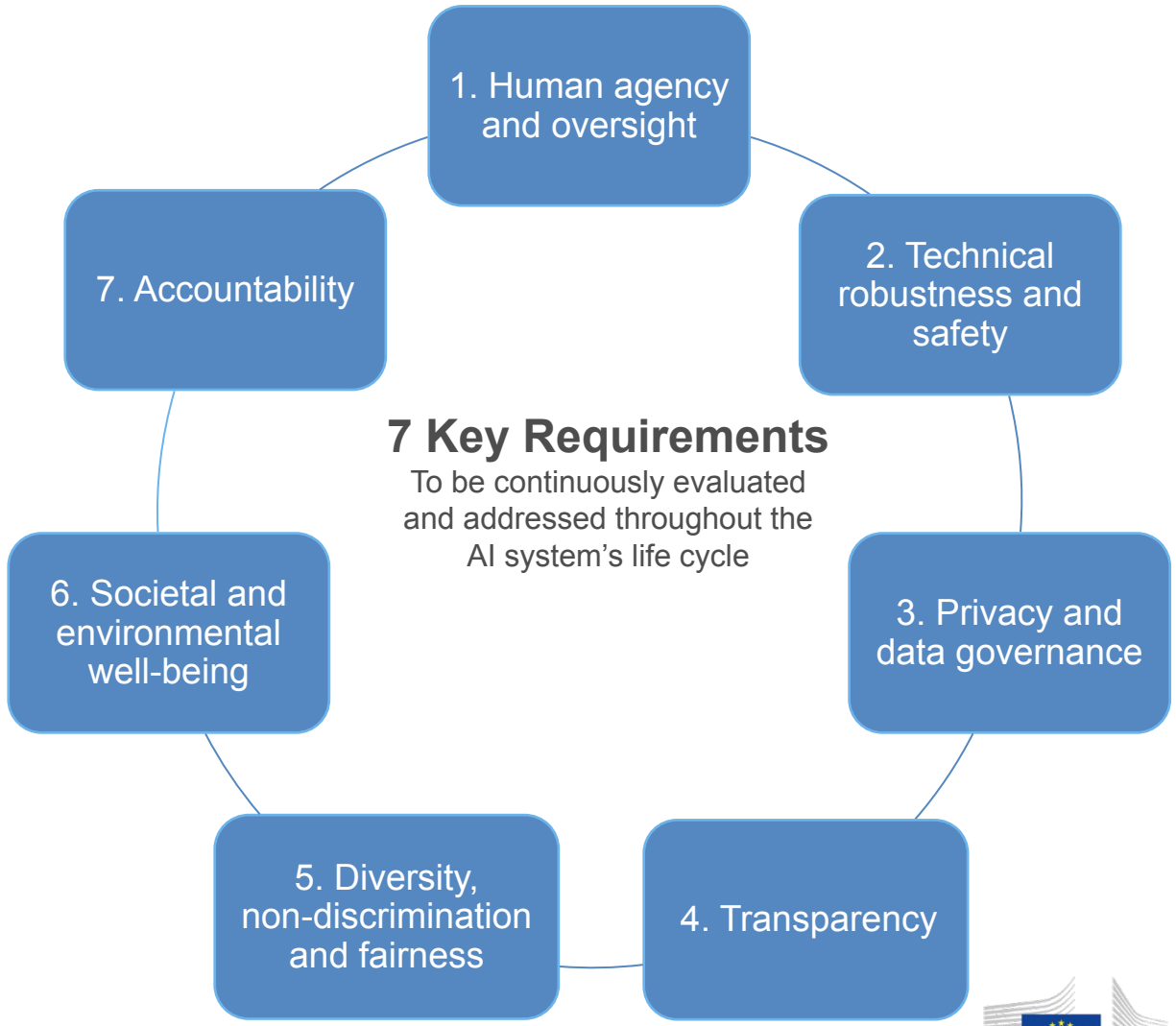
Outline

1. Intro
2. **EU approach for trustworthy AI**
3. Human behaviour and machine intelligence

HLEG on AI - 2018

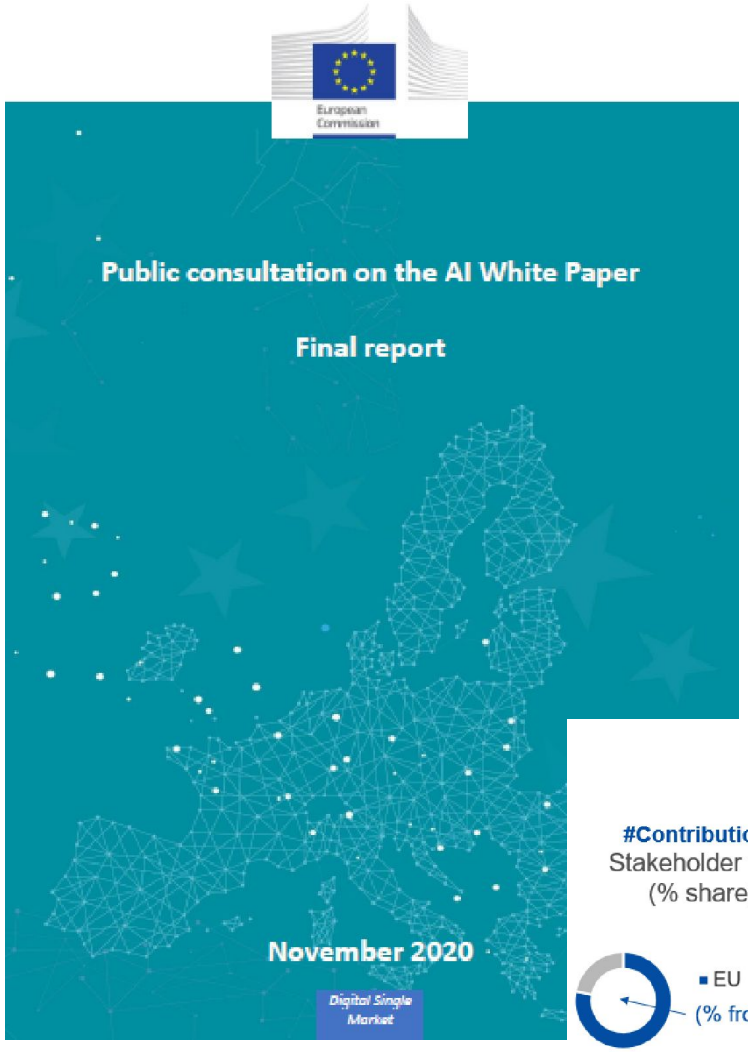
European Commission

Ethics guidelines for Trustworthy AI



European Commission

AI white paper 2020

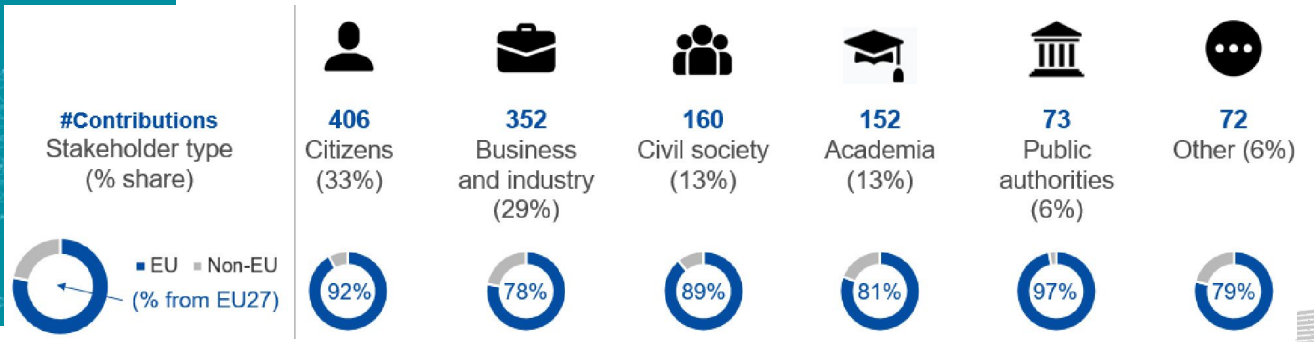


EUROPEAN COMMISSION

Brussels, 19.2.2020
COM(2020) 65 final

WHITE PAPER

On Artificial Intelligence - A European approach to excellence and trust



European Commission

European approach to AI: 2021

Communication: "*Fostering a European approach to AI*"

Ecosystem of excellence

- R&D&I
- Testing and experimentation facilities
- Digital Innovation Hubs
- Skills and talent

through

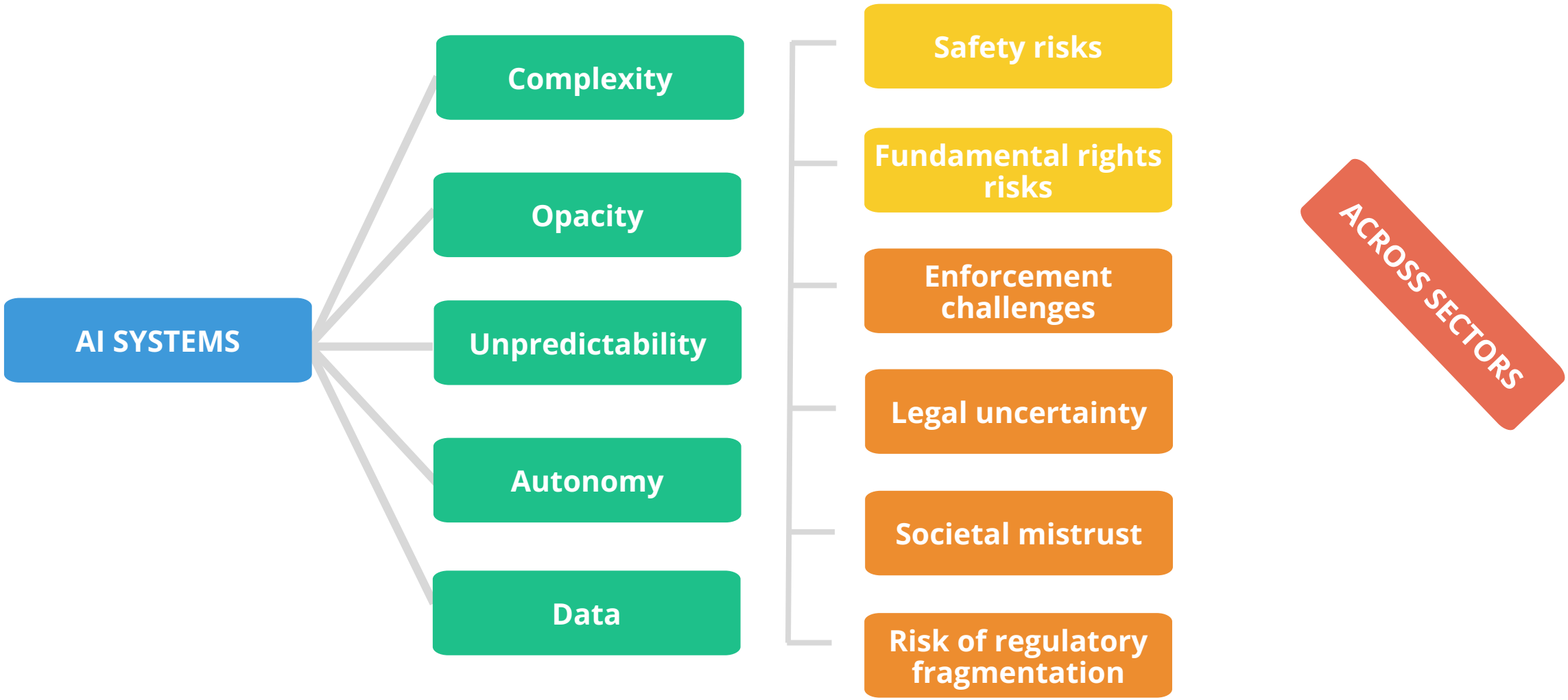
- European programmes and national activities
- Synergies in a Coordinated Plan on AI

Ecosystem of trust

- New regulatory frameworks

"..artificial intelligence will open up new worlds for us. But this world also needs rules."

Regulatory challenges



AI Act: a risk-based approach

Scope: AI system as a product.

Unacceptable risk
e.g. social scoring

Prohibited

High risk
e.g. recruitment, medical devices

Permitted subject to compliance with AI requirements and ex-ante conformity assessment

*Not mutually exclusive

'Transparency' risk
'Impersonation' (bots)

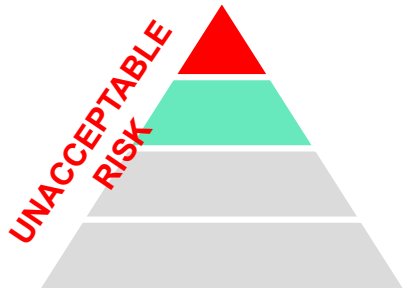
Permitted but subject to information/transparency obligations

Minimal or no risk

Permitted with no restrictions

European Commission

AI that contradicts EU values is prohibited (Title II, Art. 5)



X **Subliminal manipulation** resulting in physical/ psychological harm

X **Exploitation of vulnerabilities** resulting in physical/psychological harm

X **'Social scoring'** by public authorities

X **'Real-time' remote biometric identification for law enforcement purposes in publicly accessible spaces** (with exceptions)

European Commission

High-risk Artificial Intelligence Systems (Title III, Chapter 1 & Annexes II and III)

1 SAFETY COMPONENTS OF REGULATED PRODUCTS

(e.g. medical devices, machinery) which are subject to third-party assessment under the relevant sectorial legislation

2 CERTAIN (STAND-ALONE) AI SYSTEMS IN THE FOLLOWING AREAS

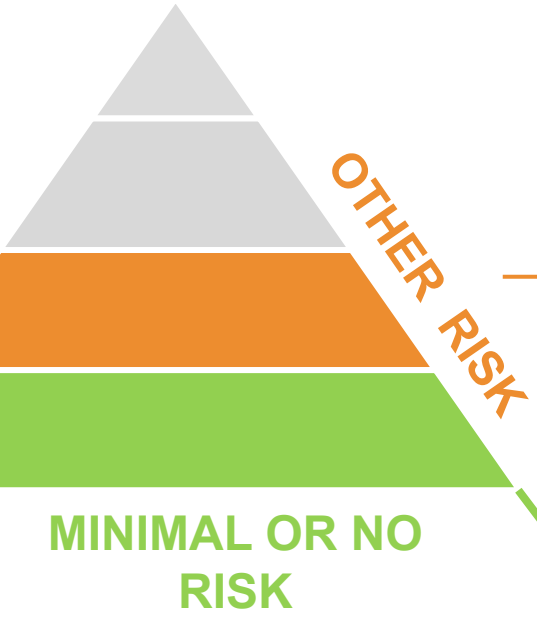
- ✔ Biometric identification and categorisation of natural persons
- ✔ Management and operation of critical infrastructure
- ✔ Education and vocational training
- ✔ Employment and workers management, access to self-employment
- ✔ Access to and enjoyment of essential private services and public services and benefits
- ✔ Law enforcement
- ✔ Migration, asylum and border control management
- ✔ Administration of justice and democratic processes

Requirements for high-risk AI systems (Title III, Chapter 2)

Establish and implement **risk management system** & in light of the **intended purpose** of the AI system

- Use high-quality **training, validation and testing data** (relevant, representative etc.)
- Draw up **technical documentation** & set up **logging capabilities** (traceability & auditability)
- Ensure appropriate degree of **transparency** and provide users with **information** on capabilities and limitations of the system & how to use it
- Ensure **human oversight** (measures built into the system and/or to be implemented by users)
- Ensure **robustness**, **accuracy** and **cybersecurity**

Most AI systems will not be high-risk (Titles IV, IX)



Transparency obligations for certain AI systems (Art. 52)

- **Notify humans** that they are **interacting with an AI system** unless this is evident
- **Notify humans** that they are **exposed to emotional recognition or biometric categorisation systems**
- Apply **label to deep fakes**

Possible voluntary codes of conduct (Art. 69)

- No mandatory obligations
- Commission and Board to encourage drawing up of codes of conduct (**voluntary application of requirements for high-risk AI systems or other requirements)**

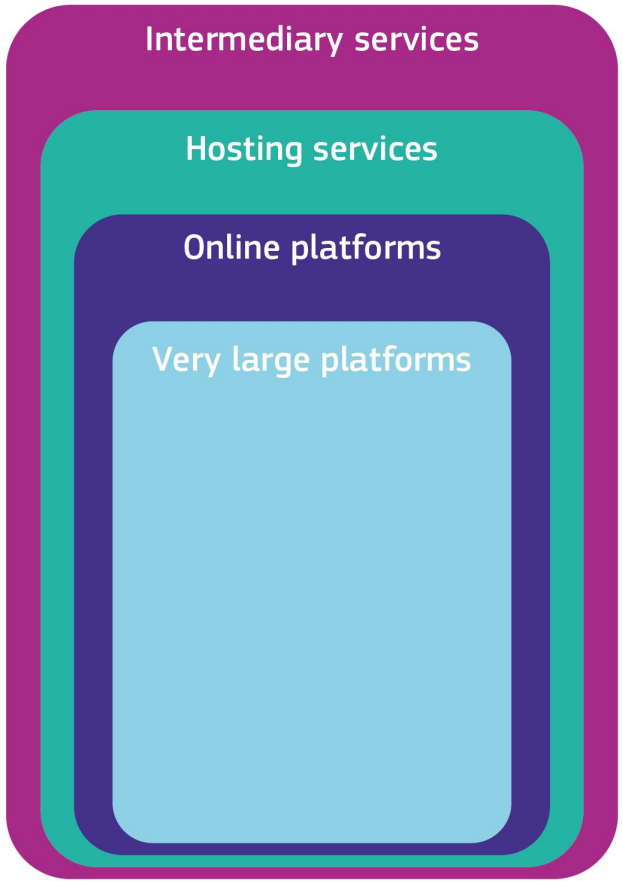
AI Act: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206
Under negotiation

European Commission

Digital Services Act: a size-based approach

Scope: digital services powered by algorithmic systems for search & recommendation

- Risk management.
- **Transparency** of recommender systems, online advertisement.
- External & independent **auditing**, internal compliance function and public **accountability**.
- **Data sharing** with authorities and researchers.
- Crisis response cooperation.



https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en

Entering into force in 2023

European Centre for Algorithmic Transparency

- Provide **technical and scientific support** to the enforcement role of the EC towards Very Large Online Platforms and Search Engines in the DSA.
- **Combine methodologies** from different disciplines.
- **Engage** with the international community of researchers and practitioners.

VLOPs:

• Alibaba Aliexpress	• Amazon Store	• Apple AppStore	• Booking.com	• Facebook	• Google Maps
• Google Play	• Google Shopping	• Instagram	• LinkedIn	• Pinterest	• Snapchat
• TikTok	• Twitter	• Wikipedia	• Youtube	• Zalando	

VLOSEs:

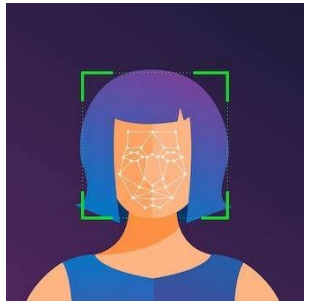
- Bing
- Google Search

https://algorithmic-transparency.ec.europa.eu
https://digital-strategy.ec.europa.eu/en/policies/dsa-vlops

Outline

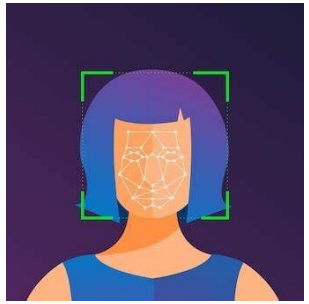
1. Intro
2. EU approach for trustworthy AI
3. **Human behaviour and machine intelligence**

Scenarios



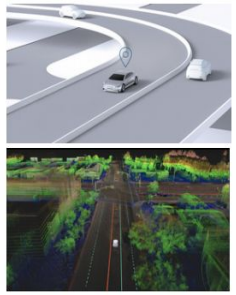
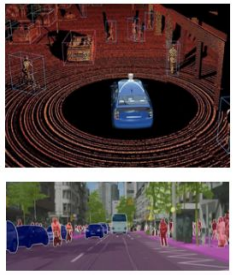
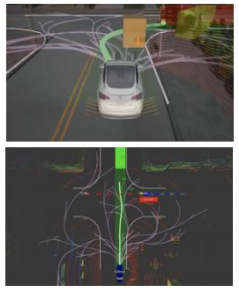
Trustworthy AI

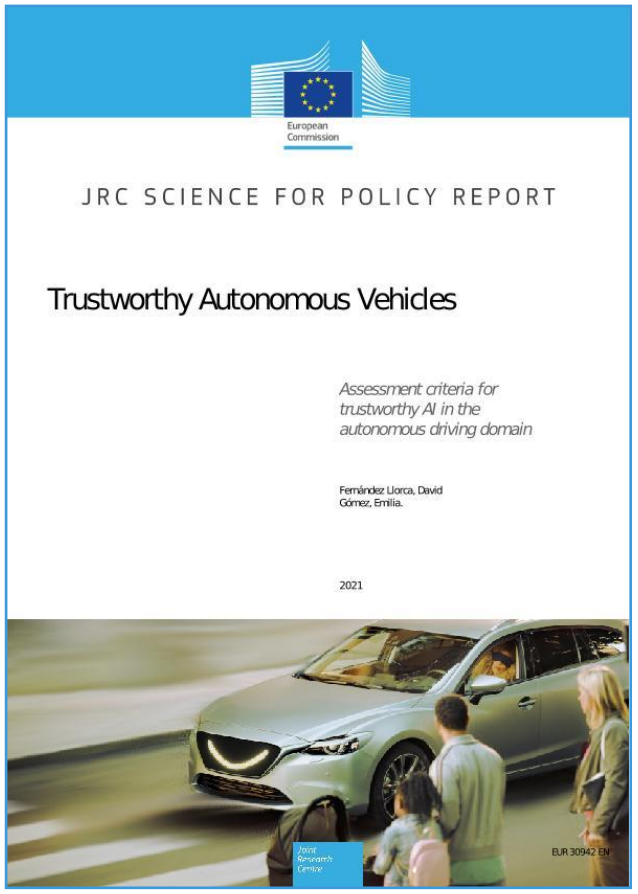
Scenarios



Trustworthy AI

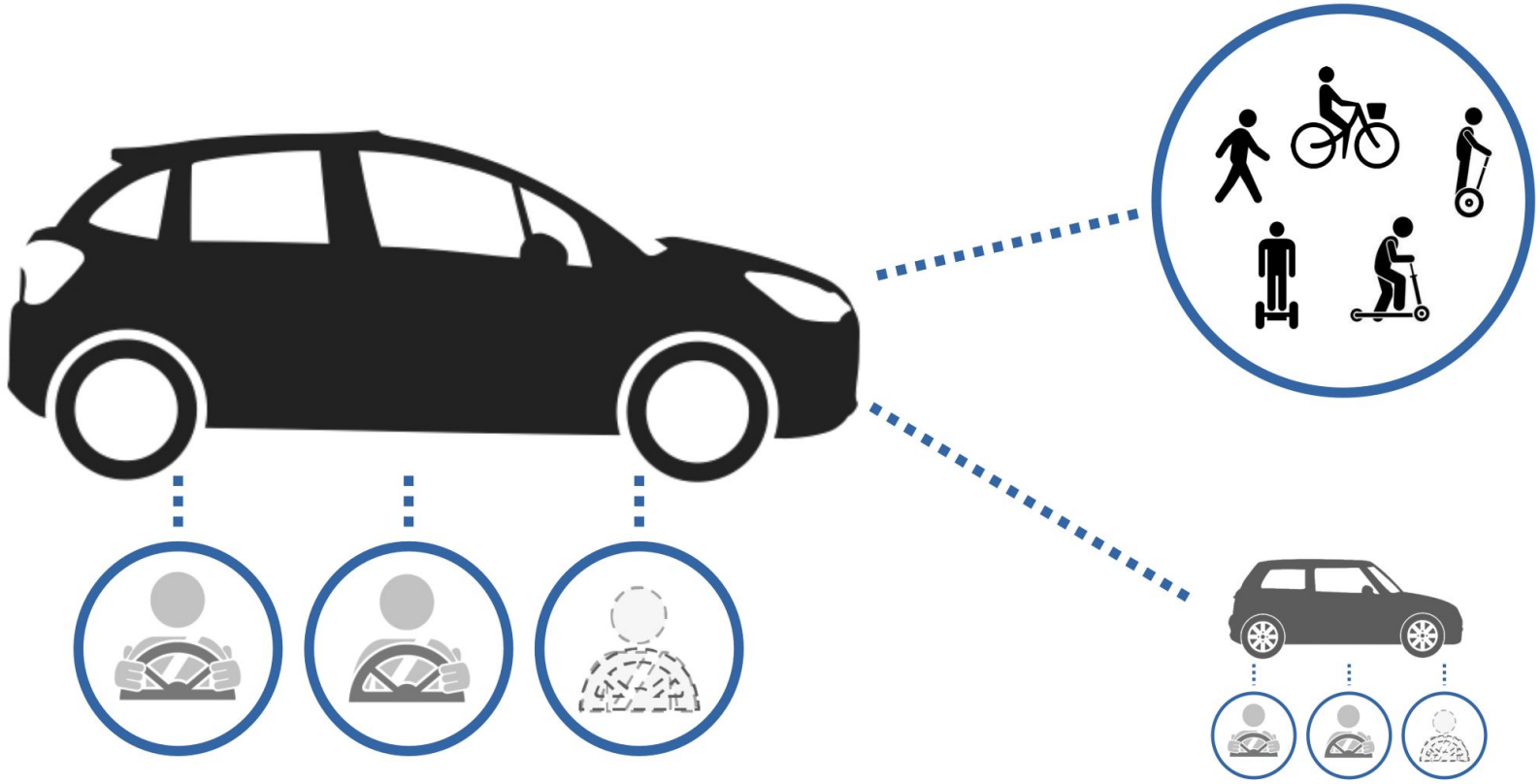
Autonomous vehicles

Localization	Dynamic Scene Understanding	Path planning	Control	User Interaction
				
				
Where am I? *Where am I heading?* *HD maps (static objects)*	*Dynamic agents* *Action/trajectory prediction* *Free space*	*Trajectories* *Environment* *Future agents' trajectories (predictions)*	*Steering* *Speed* *Lateral & Longitudinal Control*	*User commands* *Request to intervene* *Other agents' interaction*



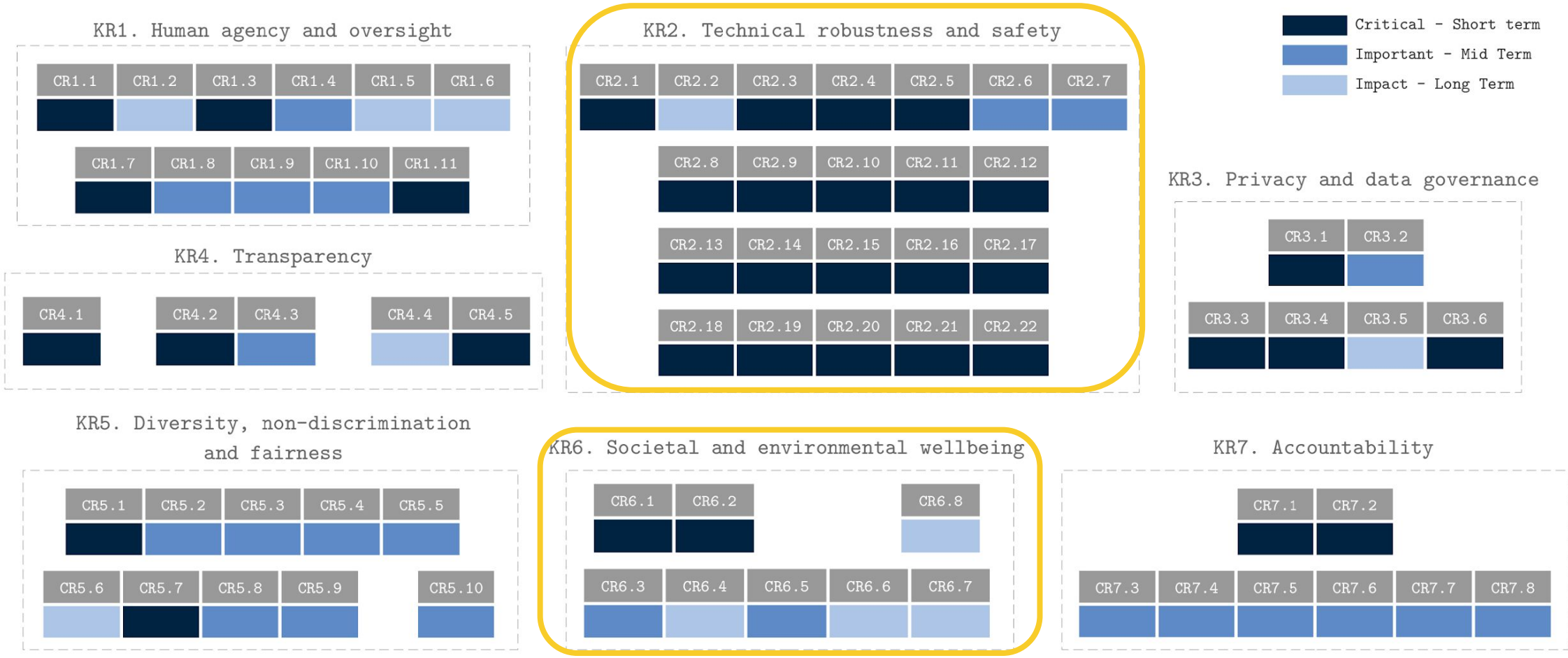
Fernandez Llorca, D. and Gomez, E., Trustworthy Autonomous Vehicles, EUR 30942 EN, Publications Office of the European Union, Luxembourg, 2021, ISBN 978-92-76-46055-8 (online), doi:10.2760/120385 (online), JRC127051.

Trustworthy Autonomous Vehicles, for whom?



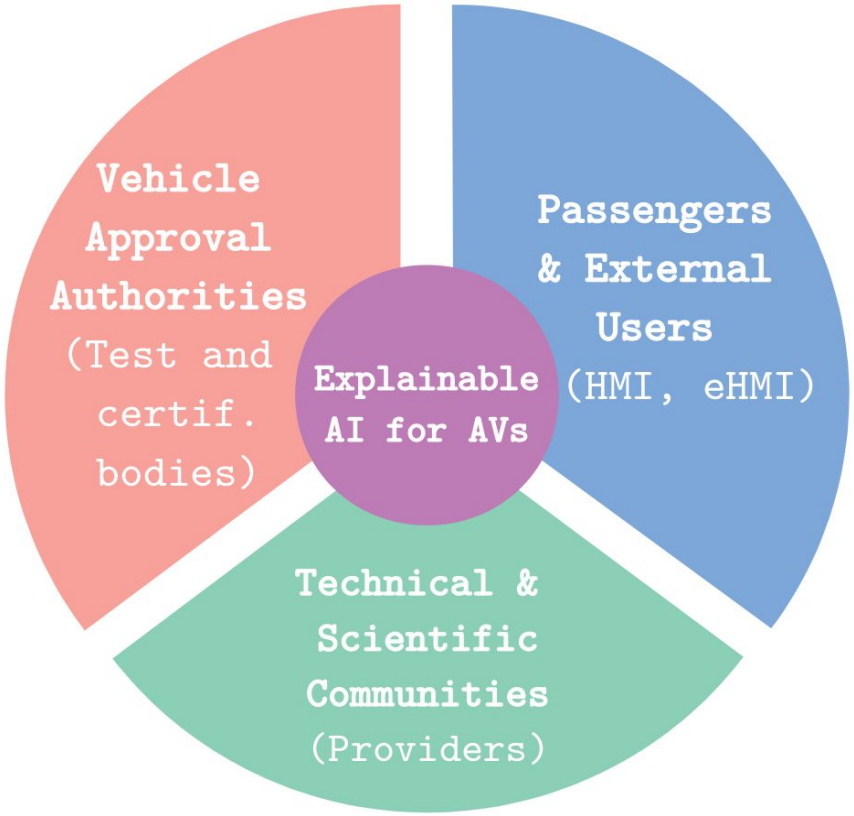
Fernandez Llorca, D. and Gomez, E., Trustworthy Autonomous Vehicles, EUR 30942 EN, Publications Office of the European Union, Luxembourg, 2021, ISBN 978-92-76-46055-8 (online), doi:10.2760/120385 (online), JRC127051.

Assessment list



Fernandez Llorca, D. and Gomez, E., Trustworthy Autonomous Vehicles, EUR 30942 EN, Publications Office of the European Union, Luxembourg, 2021, ISBN 978-92-76-46055-8 (online), doi:10.2760/120385 (online), JRC127051.

Transparency

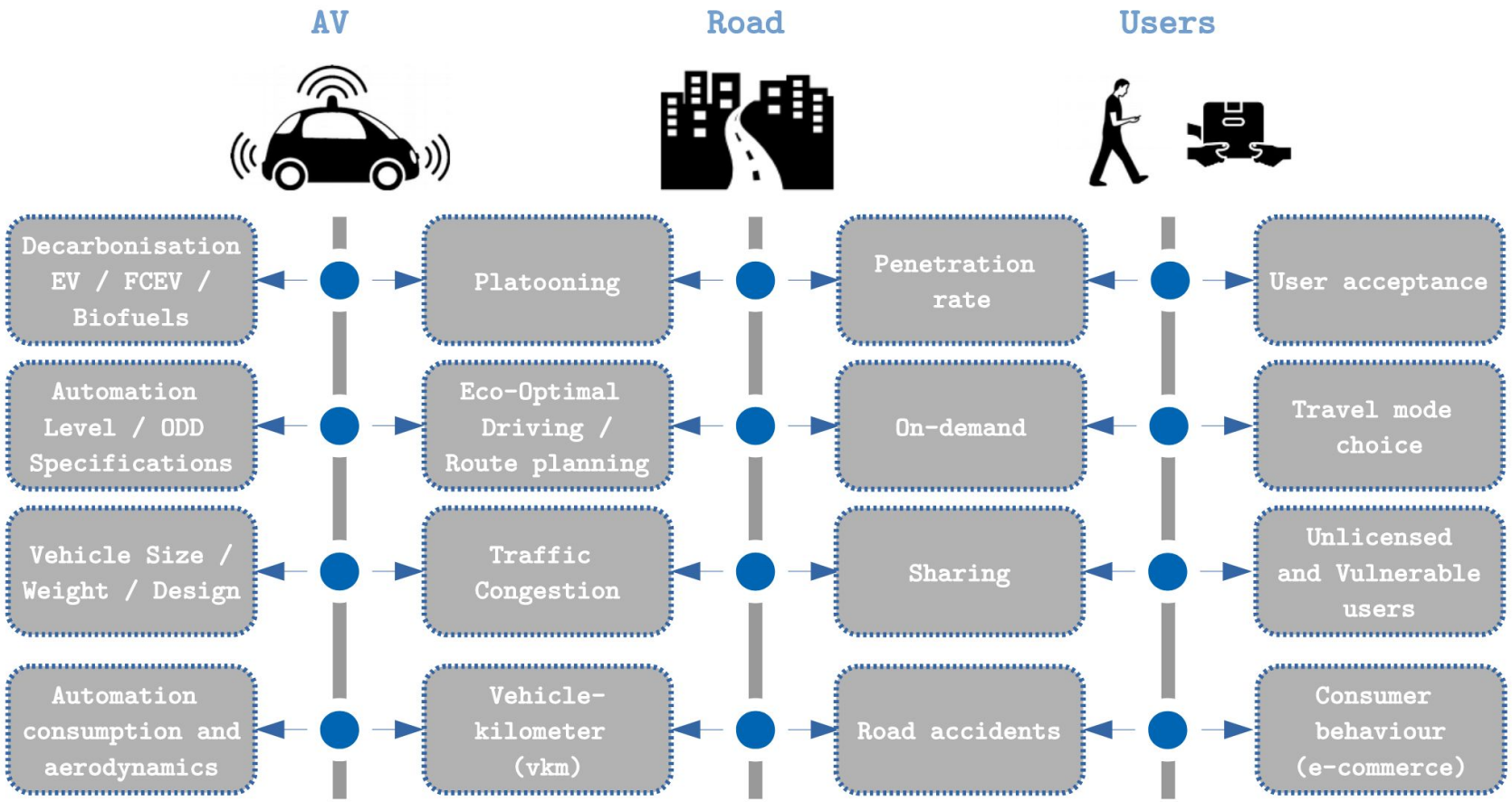


- AI traceability and adequate **logging practices**.
- Explainability barriers/questions for the different components.

AV Layers	Explainability Barriers	Explainability Questions
Localization	- Multiple sensors types - Fusion of multiple systems - Map-reality gap & Driver	- Is localization accuracy enough? - How close or far are we from exiting or entering a pre-mapped region (e.g. ODD)? - How will the localization system behave in unmapped scenarios? - Is localization fail-x (aware, safe and operational)?

Fernandez Llorca, D. and Gomez, E., Trustworthy Autonomous Vehicles, EUR 30942 EN, Publications Office of the European Union, Luxembourg, 2021, ISBN 978-92-76-46055-8 (online), doi:10.2760/120385 (online), JRC127051.

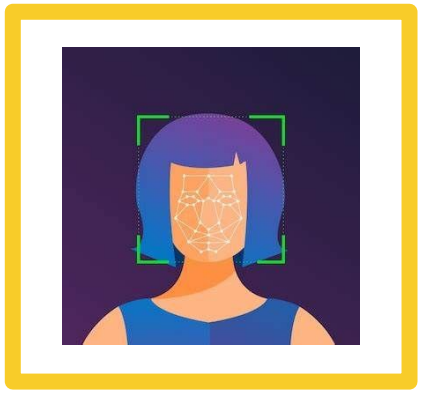
Societal and environmental well-being



Fernandez Llorca, D. and Gomez, E., Trustworthy Autonomous Vehicles, EUR 30942 EN, Publications Office of the European Union, Luxembourg, 2021, ISBN 978-92-76-46055-8 (online), doi:10.2760/120385 (online), JRC127051.

European Commission

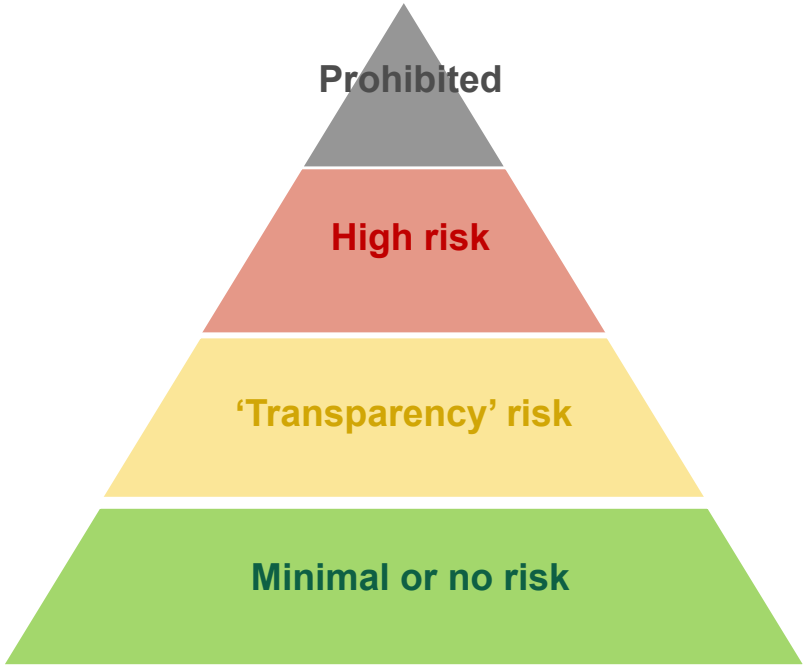
Scenarios



Trustworthy AI

Establishing the landscape of facial processing

- > 37K scientific publications.
- 183 companies.
- 60 real-world applications.
- Application areas, risk level (AI Act proposal), academic references and key companies.



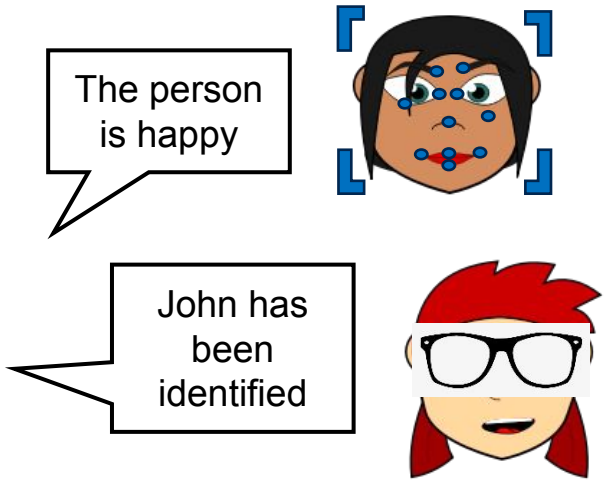
Hupont, I., Tolan, S., Gunes, H. and Gómez, E. The landscape of facial processing applications in the context of the European AI Act and the development of trustworthy systems. *Sci Rep* **12**, 10688 (2022). https://doi.org/10.1038/s41598-022-14981-6

Computational tasks

Input:
Facial images or videos

FACIAL PROCESSING

Face detection	Facial landmak extraction	Face tracking
Face verification	Face recognition	Kinship verification
Face spoofing detection	Facial expression recognition	AU detection
Facial attribute estimation	Facial attribute manipulation	Automatic lip reading



Output:
High-level information or processed image

Hupont, I., Tolan, S., Gunes, H. and Gómez, E. The landscape of facial processing applications in the context of the European AI Act and the development of trustworthy systems. *Sci Rep* **12**, 10688 (2022). https://doi.org/10.1038/s41598-022-14981-6

Facial processing applications

ID	Risk	Application	Computational tasks	Areas	# Companies	
					SME	Large
BI1	●●	Access control	FD + FI (+FSD)	BIC, MCI, EDU, EMP, LE, VSU, TRA, ENT, CLI, TOU, FIN, IND	33	20
BI2	●●	Access control with masks	FD (+FAM) + FI (+FSD)	idem	4	6
BI3	●	Border control*	FD + FV + FSD	BIC, MIG, LE	6	10
BI4	●	Banking authentication*	FD + FV + FSD	BIC, FIN, MKT	11	13
BI5	●●	Sousveillance (video surveillance at human level using, e.g., bodycams)	FD (+FT) + FI	BIC, LE, MCI, VSU	9	1
BI6	●	Devices, machines and data unlocking*	FD + FV + FSD	BIC, MCI, ENT, IND, TRA, CLI	13	14
BI7	●	Face authentication for e-Government*	FD + FV	BIC, SER, JUS, EMP, POL, CLI	1	5
BI8	●●	Unconstrained face identification	FD (+FT) + FI	BIC, LE, MIG, MCI, VSU	33	14
BI9	●●	Person re-identification	FD + FT + FI	BIC, LE, MCI, VSU	3	3
BI10	●	Person search by identity†	FD (+FT) + FV	BIC, LE, VSU, ENT	23	8
BI11	●●	Contact tracing†	FD + FT + FI	BIC, LE, CLI	4	0
BI12	●●	Person tracking with drones	FD + FT + FI	BIC, LE, VSU	2	0
BI13	●●	Perimeter protection	FD + FT + FI	BIC, LE, MCI, VSU	5	3
BI14	●	Control of attendance	FD + FV/FI	BIC, EMP, EDU	17	9
BI15	●	VIP recognition	FD (+FT) + FI	BIC, MKT, ENT, TOU, FIN	14	1
BI16	●	Face tagging in personal pictures and videos	FD + FI	BIC, ENT	3	9
BI17	●	Assistance for people with visual impairments	FD (+FT) + FI (+FER)	BIC, SOC, CLI	0	1
BI18	●●	Person search in social networks†	FD + FV	BIC, LE, EMP, SER, MKT, POL	1	0
BI19	●●	Mobile surveillance robots	FD (+FT) + FI	BIC, LE, MCI, VSU, IND	2	0
BI20	●	Product personalisation	FD (+FT) + FI	BIC, ENT, TRA, MKT	2	3
BC1	●	Demographic analysis	FD + FT + FAE	BIC, MKT, TOU	21	9
BC2	●●	Person search by facial appearance	FD (+FT) + FAE	BIC, LE, VSU, ENT	1	1
BC3	●●	Face mask detection	FD (+FT) + FAE	BIC, LE, CLI, VSU, TOU, MKT, TRA	13	6
BC4	●	Decision-making based on detected personal attributes	FAE	BIC, EDU, EMP, SER, MIG, JUST, SOC, FIN	0	0
BC5	●	Personalisation of advertising content	FD + FAE	BIC, MKT	4	0
BC6	●	Verification for age-restricted goods	FD + FAE	BIC, MKT, ENT	2	1
BC7	●	Clinical syndrome assessment	FD (+FT) + AU/FAE/FER	BIC, CLI	1	0

Trustworthy AI in facial processing

Fairness and **datasets:**

- Demographically imbalanced.
- Big Techs vs SMEs.

Software architectures tend to be increasingly distributed:

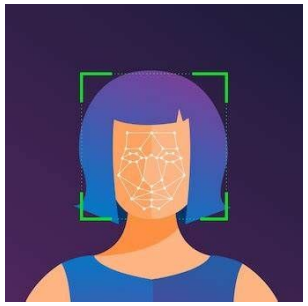
- Security and **privacy** issues.
- Federated learning, visual cryptography, data minimisation…

Need for **evaluation benchmarks**

- Neglected factors: energy consumption, fairness, explainability, human oversight.
- Operational settings, intended purpose.

Hupont, I., Tolan, S., Gunes, H. and Gómez, E. The landscape of facial processing applications in the context of the European AI Act and the development of trustworthy systems. *Sci Rep* **12**, 10688 (2022). https://doi.org/10.1038/s41598-022-14981-6

Scenarios



Trustworthy AI

Our work

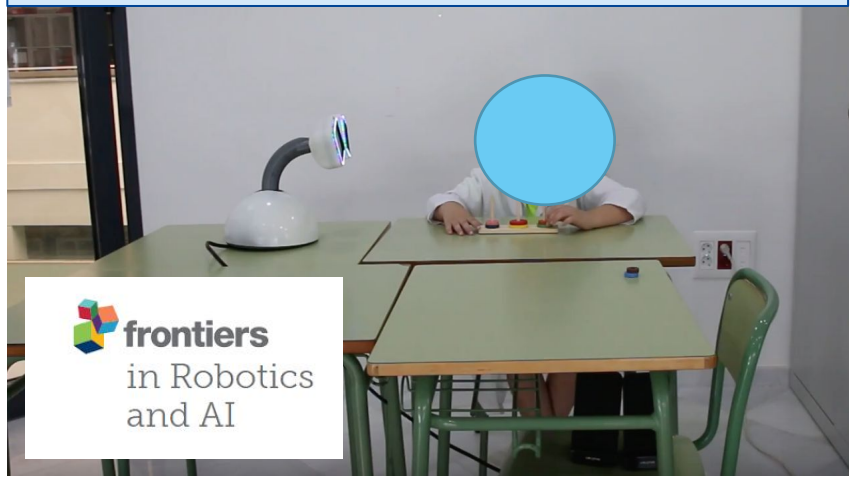
Pic by Wayan Vota flickr.com/photos/dcmetroblogger/6574651159

https://spectrum.ieee.org/honda-research-institute-haru-social-robot

- Field studies
- Mapping
- Design & evaluation
- Science for policy

Field studies: social robots

(1) Impact of a social robot on child's cognitive processes in a problem-solving task (20 children)



(2) Impact of the social positioning of a robot on child-child social interaction (84 children)

(3) Children perception and trust (84 children)

Charisi, V., Gomez, E., Mier, G., Merino, L., & Gomez, R. (2020). Child-Robot Collaborative Problem-Solving and the Importance of Child's Voluntary Interaction: A Developmental Perspective. *Frontiers in Robotics and AI*, *7*, 15.
Charisi V., Merino, L., Caballero, F., Escobar, M., Gomez, R., Gomez, E. The Effects of Robot Cognitive Reliability and Social Positioning on Child-Robot Team Dynamics. International Conference on Robotics and Automation (ICRA2021).
Escobar, M., Charisi, V., Gómez, E. I've seen a robot!: The impact of cognitive reliability and expressivity in children's perception of a robot. CSCW

European Commission

Mapping: recommender systems

Pic by Wayan Vota flickr.com/photos/dcmetroblogger/6574651159

Opportunities

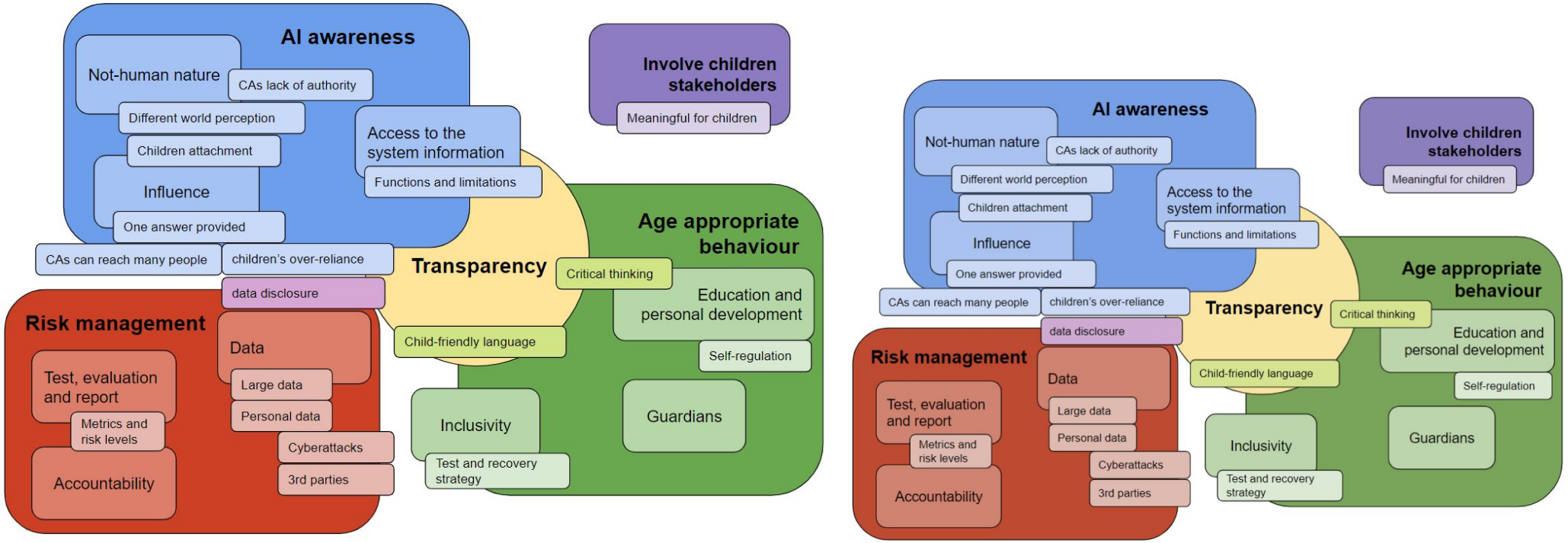
- Bring value and support children´s autonomy
- Accessibility to material on a large-scale
- Self-guided, personalised learning
- Interaction and peer-to-peer recommendation
- Areas: information search, video rec., music rec., learning, smart toys, story and book rec., social media.

Risks

- Privacy
- Over-exposure, information bubbles
- Undesirable content
- Advertising
- Addictions or dependency
- Difficulty for parents to monitor children's behaviour
- Propagation of certain stereotypes (e.g. gender)

Gómez, E., Charisi, V., and Chaudron, S. Evaluating recommender systems with and for children: towards a multi-perspective framework, PERSPECTIVES workshops, Recsys 2021.

Design & evaluation: conversational agents

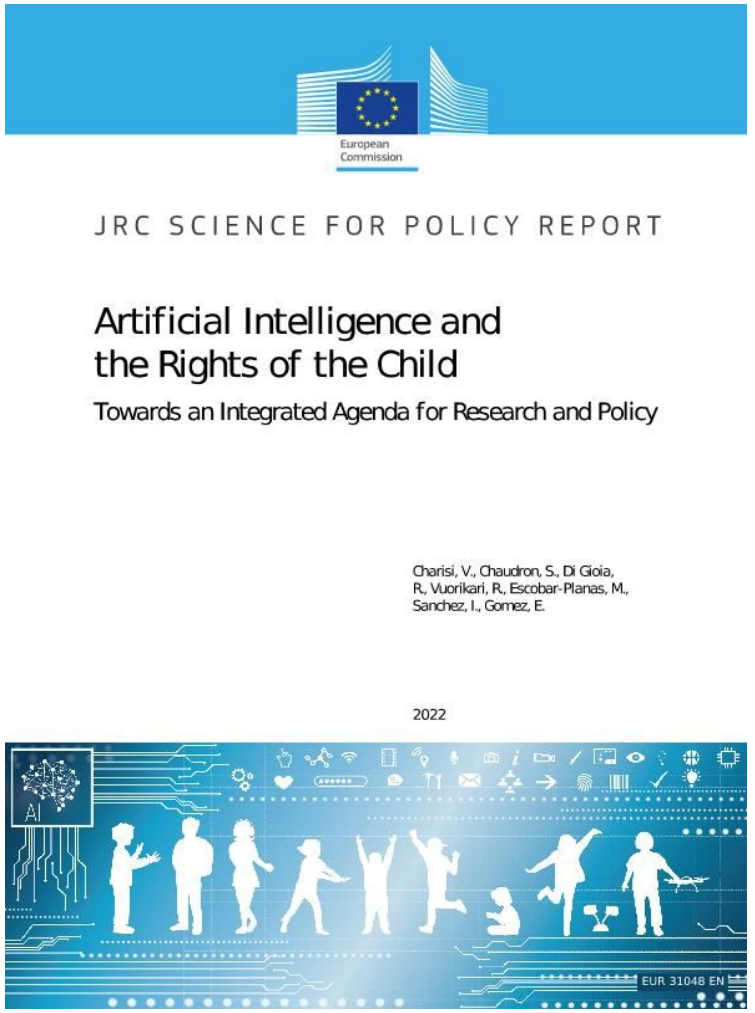


Escobar-Planas, M. 2022. Towards Trustworthy Conversational Agents for Children. In Interaction Design and Children (IDC '22). ACM, 693–695. https://doi.org/10.1145/3501712.3538826

Escobar-Planas, M., Gómez, E., Martínez-Hinarejos, C. Guidelines to Develop Trustworthy Conversational Agents for Children, Ethicomp, 2022. https://arxiv.org/abs/2209.02403

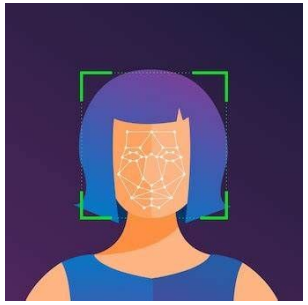
Science for policy: 5 key findings

1. Make strategic and systemic choices.
2. Child-friendly transparency measures.
3. Need for comprehensive studies.
4. Multi-perspective evaluation.
5. Children cognitive stage adaptation.



https://publications.jrc.ec.europa.eu/repository/handle/JRC127564

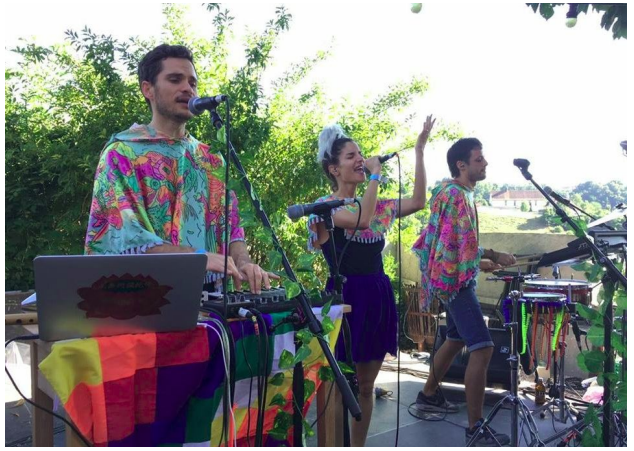
Scenarios



Trustworthy AI

Trustworthy AI & music

- Considered with low/minimal risk.
- Link to culture, emotions, creativity.

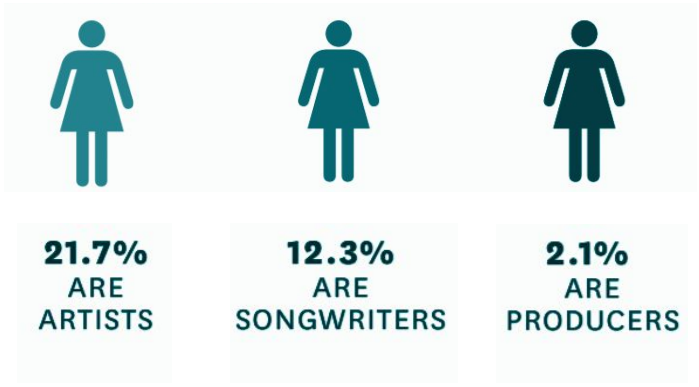


From left to right: Shibusashirazu Orchestra, The Cambodian Space Project, Sun Ra, Chancha Via Circuito

upf. Universitat Pompeu Fabra Barcelona — MTG Music Technology Group

European Commission

Fairness: gender bias



Pre-existing Gender Bias - data

Strong pre-existing bias towards male artists on the Last.FM platform.

Gender Bias Propagation - algorithm

Pre-existing bias drive Collaborative Filtering-based algorithms to over-represent male artists

Differences Across Algorithmic Approaches

Model-based approach produces recommendations more representative of user's input gender preference vs memory-based approachs.

Dougal Shakespeare, Lorenzo Porcaro, Emilia Gómez, and Carlos Castillo. (2020) *Exploring Artist Gender Bias in Music Recommendation*. 2nd Workshop on the Impact of Recommender Systems (ImpactRS), co-located with the 14th ACM Conference on Recommender Systems (RecSys 2020). Virtual, 22nd-26th September (pdf).

Transparency

MOTHERBOARD

‘Deep Voice’ Software Can Clone Anyone's Voice With Just 3.7 Seconds of Audio

Using snippets of voices, Baidu's ‘Deep Voice’ can generate new speech, accents, and tones.

SHARE TWEET

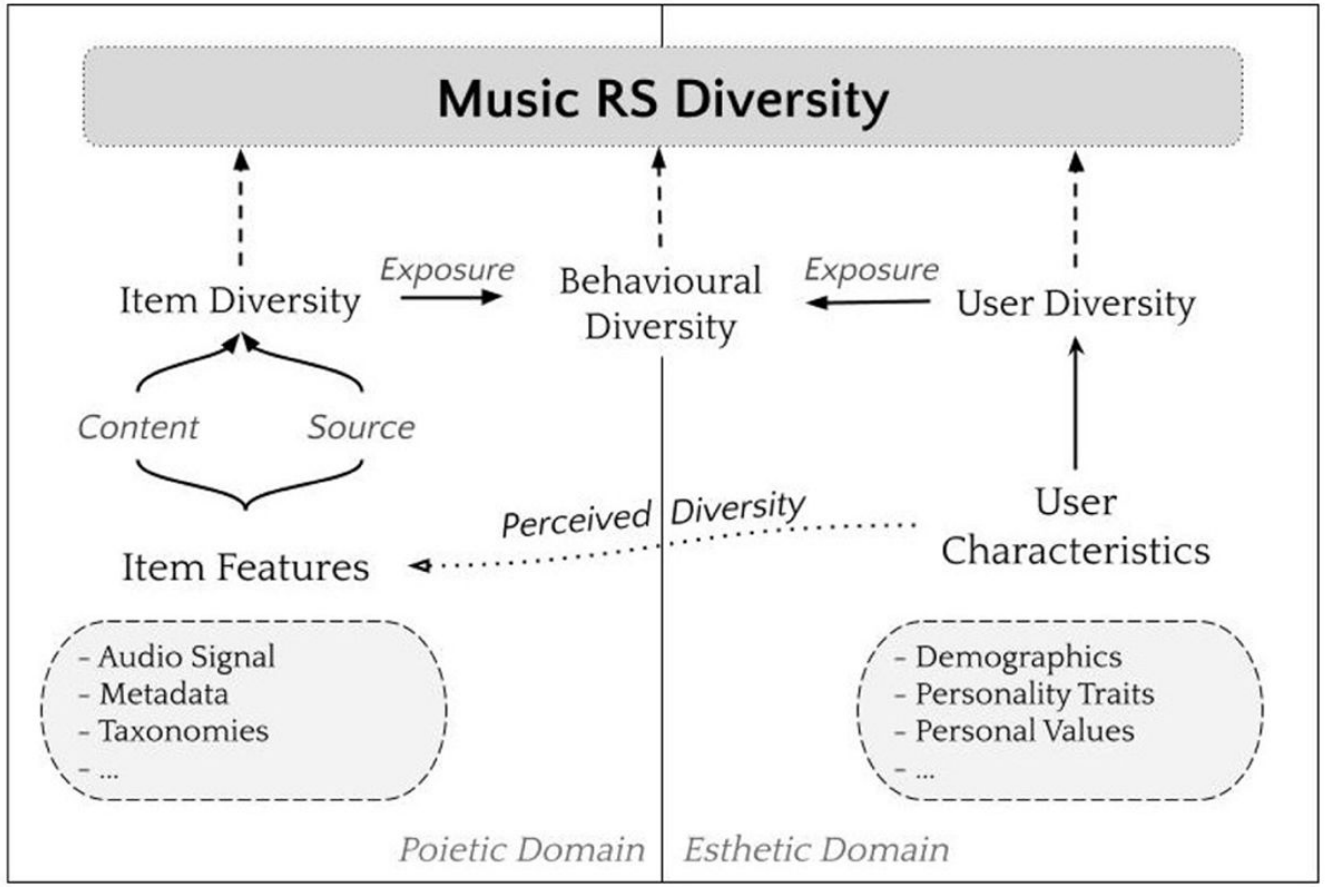
- Towards listeners.
- Towards creators, intellectual property.

Gómez, E., Blaauw, M., Bonada, J., Chandna, P., & Cuesta, H. (2018). Deep learning for singing processing: Achievements, challenges and impact on singers and listeners. arXiv preprint arXiv:1807.03046.

Sturm BLT, Iglesias M, Ben-Tal O, Miron M, Gómez E. Artificial Intelligence and Music: Open Questions of Copyright Law and Engineering Praxis. *Arts*. 2019; 8(3):115. https://doi.org/10.3390/arts8030115

European Commission

Diversity-by-design in music Recsys



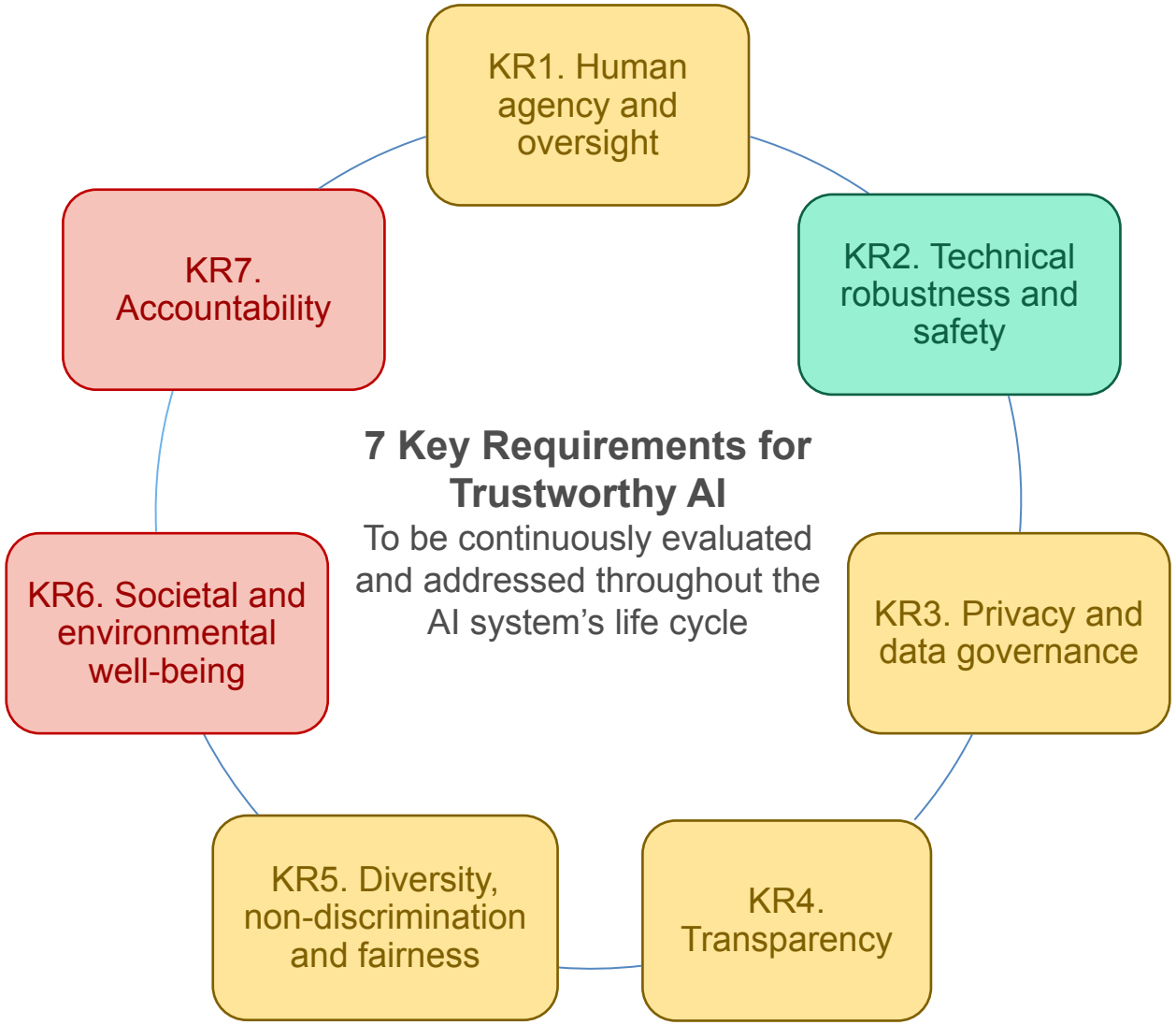
Porcaro, L., Castillo, C. and Gómez, E., 2021. Diversity by Design in Music Recommender Systems. *Transactions of the International Society for Music Information Retrieval*, 4(1), pp.114–126. DOI: http://doi.org/10.5334/tismir.106

Conclusions

Newly addressed

Some research

Strong background



European Commission

Scientific challenges, practical methodologies and policy perspectives for trustworthy AI

Emilia Gómez (emilia.gomez-gutierrez@ec.europa.eu)

Work with the HUMAINT team

https://ai-watch.ec.europa.eu/humaint

European Commission