European Commission

SCIENCE FOR POLICY BRIEF

Trustworthy AI and Automated Driving

Artificial Intelligence in Autonomous Vehicles: towards trustworthy systems

HIGHLIGHTS

→ As Artificial Intelligence (AI) is the main enabler of Autonomous Vehicles (AVs), and autonomous mobility is a scenario of high-risk nature, future sectorial regulations of AVs are expected to be aligned with the AI Act.

→ Beyond requirements of safety and robustness, other important criteria to be considered include human agency and oversight, security, privacy, data governance, transparency, explainability, diversity, fairness, social and environmental well-being and accountability.

→ These trustworthy requirements for AVs have a heterogeneous level of maturity and bring new research and development challenges in different areas. A specific analysis of the evaluation criteria for trustworthy AI in the context of autonomous driving is needed.

→ There is a window of opportunity to define a European approach to AVs in future implementing acts, by including requirements of trustworthy AI systems in harmonized procedures for the type-approval of AVs at EU level.

'A European approach to future harmonized sectorial procedures for the type-approval of autonomous vehicles can be based on the requirements of trustworthy artificial intelligence'

FROM TRUSTWORTHY ARTIFICIAL INTELLIGENCE TO TRUSTWORTHY AUTONOMOUS VEHICLES

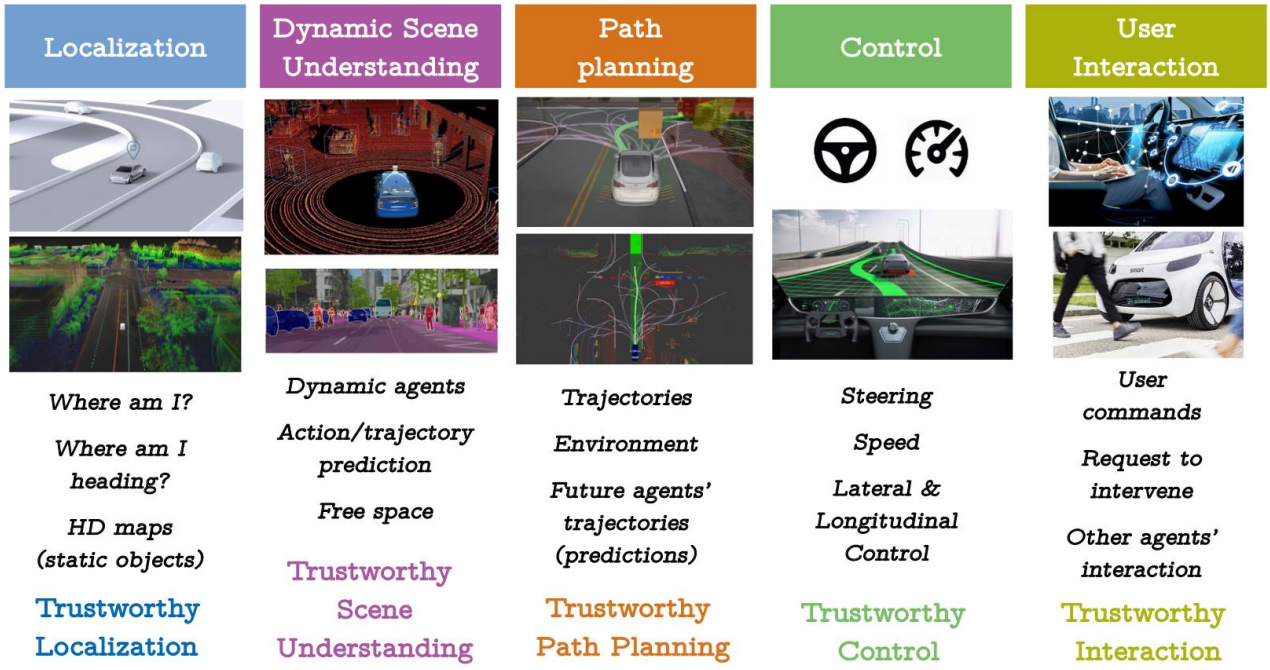
Autonomous Vehicles and Artificial Intelligence

It is no coincidence that advances in autonomous driving are developing in parallel with those in AI as **the main enabler for *assisted*, *automated* and *autonomous*[1] driving is AI**.

[1] The approach proposed to refer to vehicles with automated driving systems is to consider *assisted* for SAE Levels 1 and 2 (driver), *automated* for SAE Level 3 (backup driver) and *autonomous* for SAE Levels 4 and 5 (passenger).

Joint Research Centre

Figure 1 – Main technology layers of an AV, each one powered by one or multiple AI systems.



Source: Trustworthy Autonomous Vehicles, 2021, EC JRC.

In fact, AVs can be seen as a set of multiple, complex and interrelated AI systems, embodied in the form of a car. The five key technology layers of AVs are **localisation**, **dynamic scene understanding, path planning, control** and **user interaction** (see Fig. 1). AI is the predominant technology in most of them, in some cases, indispensable.

Therefore, when referring to **trustworthy AI** systems for AVs, it seems acceptable to extrapolate the concept to refer to it as **trustworthy AVs**.

The term *trustworthy* should not be interpreted in its literal sense, but as a global **framework** that includes multiple principles, requirements and criteria. These elements were established by the High Level Expert Group on Artificial Intelligence (AI HLEG) in the assessment list for trustworthy AI systems as a mean to maximise the benefits while minimising the risks.

The AI Act and future implementing acts for autonomous vehicles

AVs could enable **new mobility services** and car sharing schemes to respond to the increasing demand for mobility of people and goods. They could significantly **improve road safety** as human factors (errors, distractions, violations of the traffic rules) play a key role in most accidents and bring **mobility** to those **who cannot drive themselves**. In addition, they could **accelerate vehicle electrification** and **free up urban public spaces** currently used for parking.

But as with any disruptive technology, its adoption also entails some risks. The operation of AVs takes place in public spaces potentially endangering the users (occupants) and the public (external road users), affecting unknown and not identifiable persons without prior consent. They can cause severe physical harm, even death, as well as property damage. The inherently **high-risk nature of AVs** can be confidently assumed and the AI systems of AVs are clearly used as safety components.

The above suggests that the proposal for a regulation laying down harmonized rules on AI (**AI Act**, COM (2021)206) which establishes a set of requirements (to be further developed) that AI systems must meet if they are operated in high-risks scenarios, will be **highly relevant** for developing future harmonized sectorial **procedures and technical specifications for the type-approval of AVs**.

Following the approval of the next implementing act for the application of Regulation (EU) 2019/2144 with regard automated driving systems (ADS), and **as scale increases**, it is reasonable to assume that **future implementing acts** will need to be **consistent with the AI Act**.

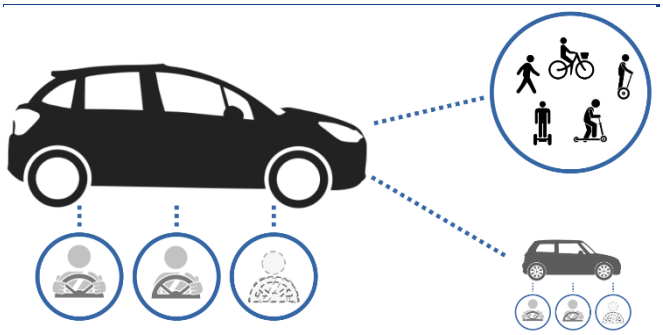
The requirements defined in the AI Act are based on a subset of the 7 requirements for **trustworthy AI** systems proposed by the AI HLEG. The detailed analysis of the application of such assessment criteria of trustworthy AI for AVs can serve as a basis to advance in a future definition of **a European approach to AVs**, **in line with the Coordinated Plan on AI, the AI Act**, the strategy on Sustainable and Smart Mobility (COM(2020) 789), and **in parallel to the work developed by UNECE (WP.29/GRVA).**

TRUSTWORTHY AUTONOMOUS VEHICLES

Trustworthy AVs, for whom?

In the field of AVs when thinking about users (i.e., **human-centric**), we need to consider multiple stakeholders: **in-vehicle users** which include backup drivers for automated vehicles or passengers for autonomous vehicles, and **external road users**, including **vulnerable** road users (e.g., pedestrians, cyclists, users of personal mobility devices), drivers of conventional vehicles and drivers/passengers of other automated/autonomous vehicles (see Fig. 2). This is a major challenge as it sometimes involves **conflicting perspectives and interests** that require compromise solutions.

Figure 2 – Interaction and communication of AVs with drivers/passengers and external road users. User-centric design should address two dimensions and multiple types of users.



Source: Trustworthy Autonomous Vehicles, 2021, EC JRC.

Using a qualitative methodology, we can, on the one hand, establish the maturity level, relevance and time horizon of each requirement for trustworthy AI systems for AVs and, on the other hand, analyse the state of the art of each of the seven requirements proposed by the AI HLEG. In what follows, the most relevant conclusions are presented for each of them.

KR1. Human agency and oversight

Human agency for AVs is linked to the principle of human autonomy, affecting acceptance (e.g., disuse) and safety (e.g., misuse). **New agency-oriented** in-vehicle and external **Human Machine Interfaces** (HMIs) are needed to ensure an adequate level of human agency. Efficient approaches to measure and calibrate the sense of agency are required.

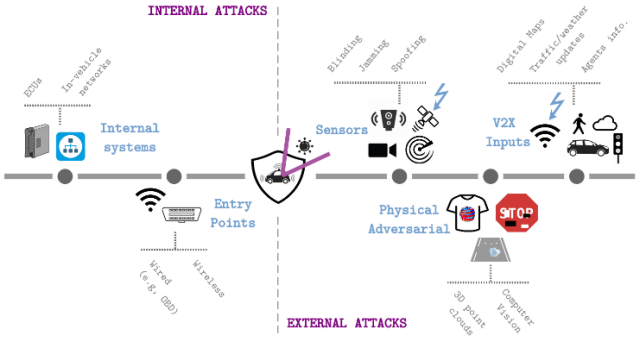
Human oversight for AVs is exercised differently depending on the level of automation. It is also exercised to some extent by external road users, with the risk of abuse in the interaction knowing that AVs will stop in any case. For **a proper interaction**, there must be **mutual awareness between the AV and the users** with whom it interacts.

How to effectively represent and communicate the operating status of the AV to users, including the request to intervene, is a key area of future research. Finally, oversight by drivers and passengers will require **new skills** both a priori and developed with exposure and use.

KR2. Technical robustness and safety

This requirement is linked to the principle of harm prevention, with a strong impact on user acceptance. **Attack** (see Fig. 3 for a taxonomy of attacks) **resilience and security of AVs must be addressed from a heterogeneous, constantly updated approach**, starting from security by design, including multiple defensive measures (e.g., cryptographic methods, intrusion and anomaly detection), countermeasures against adversarial attacks (e.g., redundancy, hardening against adversarial examples), fault-tolerant, fail-x, and self-healing methods, and user training.

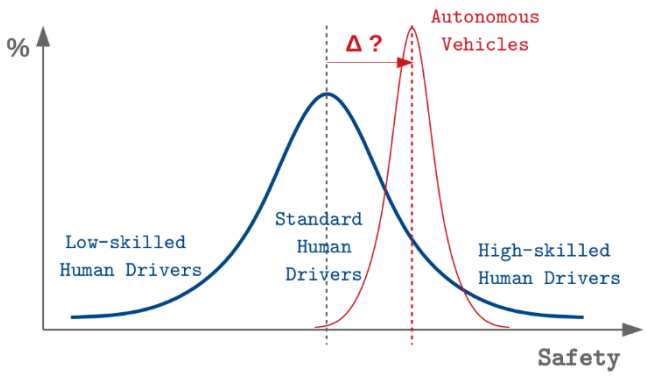
Figure 3 – Taxonomy of internal and external attacks to autonomous vehicles, including physical adversarial attacks.



Source: Trustworthy Autonomous Vehicles, 2021, EC JRC.

New innovative methods are also needed to assess the **safety of AVs against that of human drivers** that do not require endless testing periods. Expectations of safety gains (Δ in Fig. 4) that are too high could be detrimental to user acceptance. **Even small improvements in safety by AVs relative to human drivers can save many lives**, so public expectations must be appropriately calibrated so as **not to delay the adoption of AVs**, and with it the benefits of the technology.

Figure 4 – Safety performance distribution for human drivers, and autonomous vehicles, and safety gain.



Source: Trustworthy Autonomous Vehicles, 2021, EC JRC.

Important steps have been taken in the design **of new safety test procedures** for automated driving functions, including simulation, physical test in proving grounds, and real-world test drive. However, there are still important limitations, such as the **absence of real-behaviours,**

limited variability, and lack of scenarios to assess human agency and oversight, transparency or fairness.

New fall-back strategies are needed to achieve **minimal risk conditions**, as well as testing procedures to assess their safety and robustness.

Accuracy of AVs is a multi-dimensional problem, involving multiple metrics, levels, layers, use cases and scenarios. Defining **holistic metrics and thresholds to assess AVs** is a challenging research and policy-based problem to be addressed.

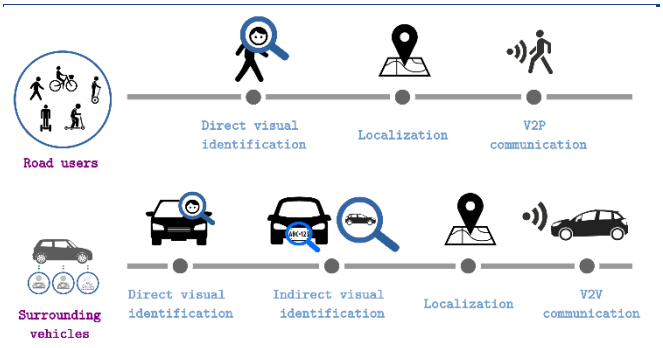
Any substantial change in an AI-based component of AVs that may modify the overall behaviour must meet all relevant trustworthy requirements and may need to be retested.

KR3. Privacy and data governance

New innovative approaches have to be implemented to **ensure data protection without negatively affecting the safety** of AVs, including agent-specific data anonymization and de-identification techniques, while preserving relevant attributes of agents.

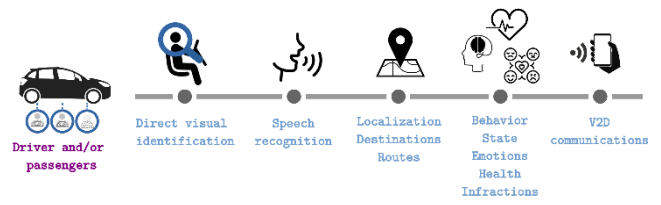
Privacy by design (also linked to security and safety) will require, among others, the encryption of data, storage devices and V2X communication channels, with a unique encryption key management system for each vehicle and including regular renewal of encryption keys.

Figure 5 – Personal data of external road users and surrounding vehicles processed by AVs.



Source: Trustworthy Autonomous Vehicles, 2021, EC JRC.

Figure 6 – Personal data of in-vehicle users (driver and/or passengers) processed by AVs.



Source: Trustworthy Autonomous Vehicles, 2021, EC JRC.

Consent to the processing of personal data (Figs. 5-6) in AVs should address two dimensions. For drivers and passengers, it should not pose any safety risk, and should include the exchange of data with other vehicles and infrastructures. For external road users, consent can be considered impossible to obtain or would involve disproportionate efforts. However, the problem can be effectively avoided if data are processed in real time or if data de-identification is properly implemented.

KR4. Transparency

Traceability is already a challenge for modern conventional vehicles, so its complexity for AVs is more than remarkable. The effective **integration** of components of **data-driven AI systems as traceable artefacts** is still an open research question.

New strategies for **intelligent data logging** must be developed to cope with the demanding requirements (bandwidth and storage capacity) of continuous data logging for AVs.

New explainable models and methods should be developed, focusing on explanations to in-vehicle and external road users, i.e. new research related to **explainable human-vehicle interaction through new HMIs and external HMIs (eHMI)**. Explainability as a requirement for vehicle type-approval frameworks will enhance the assessment of safety, human agency and oversight, and transparency, but will require new test procedures, methods and metrics.

New effective ways of **communicating** to passengers and external road users that they are **interacting with an AV** must be established, as well as new ways of communicating risks.

KR5. Diversity, non-discrimination and fairness

To avoid discrimination in decision-making, **AVs must avoid any kind of decision based on potential social values of some groups over others** (e.g., dilemmas) and must be **designed to maintain the same level of safety for all road users**. To this end, AVs may react differently to correct safety inequalities resulting from different road users' behaviours, so new real-time predictive perception and path-planning systems are needed to model the behaviour of different road users and react accordingly.

Further efforts are needed to identify possible **sources of discrimination** in state-of-the-art **perception systems** for detecting external road users according to different **inequity attributes** such as sex, age, skin tone, group behaviour, type of vehicle, colour, etc.

Unfair bias may also be present at the user-vehicle interaction layer. **Accessible and adaptable HMIs** should be designed, which is a challenge considering that AVs have the potential to extend mobility to new users.

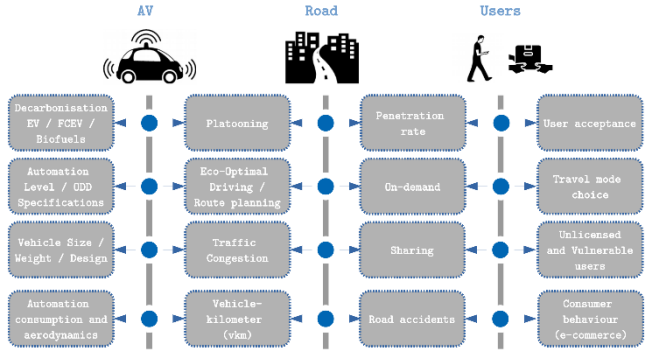
AVs opens up new autonomous mobility systems, services and products. Any **service provision approach that may discriminate against users should be avoided**.

It is necessary for **policymakers to establish a clear taxonomy of stakeholders, modulating the direction** (positive or negative) **and weight of the impact that the adoption of AVs implies for each of them**.

KR6. Societal and environmental well-being

Understanding and estimating the impact of AVs on the environment and society is a highly multidimensional and complex problem, involving many disruptive factors (see Fig. 7), for which we can only make predictions based on yet uncertain assumptions. **New approaches** and studies are needed to provide **more accurate estimates, with less uncertainty**. Policymakers must steer and monitor the adoption process to tip the balance towards a positive impact.

Figure 7 – Key environmental and social factors in the development and adoption of autonomous vehicles.



Source: Trustworthy Autonomous Vehicles, 2021, EC JRC.

Automated vehicles will not have a negative impact on jobs, but **new skills for backup drivers** will be needed. For **autonomous vehicles**, as no drivers are needed, the expected **impact** on work and skills is likely to be **negative**, but partially mitigated by the need of **non-driving tasks** less susceptible to automation and **new jobs and skills** brought by transport automation.

AVs opens up the possibility **to use travel time for work-related activities**, leading to higher productivity or a reduction of time at the workplace as commuting time could be considered as working time.

In the coming years we will see new approaches to **transform the interiors of AVs into places to work**, which is a challenge in shared mobility scenarios.

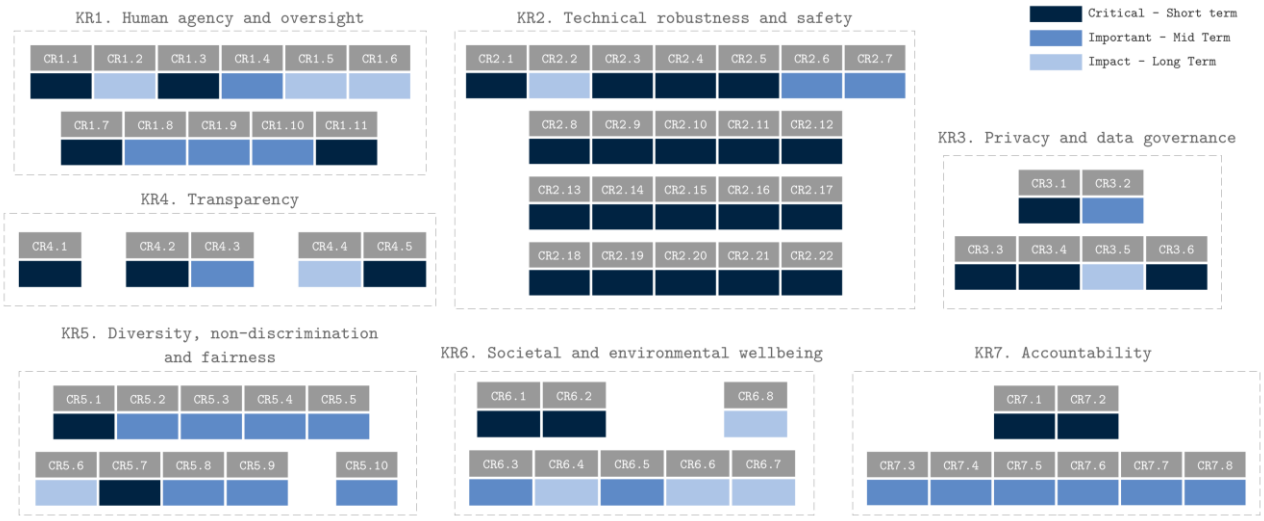
KR7. Accountability

As a safety-critical application, **AVs must be audited by independent, external auditors**. Establishing the minimum requirements for third parties to audit systems without compromising **intellectual and industrial property** then becomes a major challenge.

The same requirements and expertise needed to audit AVs would be necessary **for victims or insurers to claim for liability in accidents involving AVs**, which would be very complex and costly. **Shifting the burden of proof to the manufacturer** of AVs would make these systems more victim friendly. Considerable harmonization efforts and major updates of existing national product liability, traffic liability and fault-based liability frameworks are needed, including the Product Liability Directive and the Motor Insurance Directive.

The adoption of AVs will entail new risks, including those that are unknown at the time of production and can only emerge after-market launch. Policymakers should define **new balanced and innovative policy frameworks to accommodate insurance and liability costs between consumers** and injured parties on the one hand, **and AVs providers** on the other.

Figure 8 – Relevance and time horizon of the assessment criteria for the seven Key Requirements (KRs). Qualitative interpretation and representation based on the analysis of each of the criteria. See the source for the description of each criterion.



Source: Trustworthy Autonomous Vehicles, 2021, EC JRC.

THE WAY FORWARD

Similar to the process followed by the Commission to develop the European approach to trustworthy AI, and taking into account the relevance of AI systems in AVs and its inherent risks, it seems highly appropriate to **engage in an in-depth, multi-stakeholder discussion to define and particularise the requirements necessary for AVs to be trustworthy**.

The influence that the AI Act may have on future sectorial regulations of AVs is significant, especially as certain consistency is expected to be maintained.

Following the upcoming implementing acts for the application of Regulation (EU) 2019/2144 with regard automated driving systems (ADS), the work by UNECE (WP.29/GRVA) on AI and vehicle regulations, and the development of the AI Act (e.g., current work on the formalization and standardization of the requirements for high-risk AI systems) there is a good opportunity **to define the European approach to AVs.**

Future rules and technical procedures for the type-approval of AVs at EU level can **incorporate requirements for trustworthy systems**. That is, beyond classic requirements of safety and robustness, other important criteria can be considered such as human agency and oversight, security, privacy, data governance, transparency, explainability, diversity, fairness, social and environmental well-being, and accountability.

The application of the requirements for trustworthy AI systems for AVs involves addressing multiple problems of different nature, some of them still at a very early stage of scientific and technological maturity.

Action by policy makers to steer future regulation towards trustworthy requirements will serve as an accelerator and driver for the development and adoption of a technology that can change transport as we know it, ensuring its compliance with European values.

REFERENCES

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

.

POLICY BRIEF SERIES ON TRUSTWORTHY ARTIFICIAL INTELLIGENCE IN AUTOMATED/AUTONOMOUS DRIVING

The JRC is carrying out research on the challenges of addressing the different requirements of trustworthy artificial intelligence in the context of automated driving. This brief is one of a series of "science for policy" briefs summarizing the main outcomes on this research.

Other policy briefs published on the series includes:

Hamon, R., Junklewitz, H., Sanchez, I., Artificial Intelligence in Automated Driving: an analysis of safety and cybersecurity challenges, European Commission, 2022, JRC127189

DISCLAIMER

This policy brief was prepared by David Fernández Llorca and Emilia Gómez as part of a broader set of activities conducted within the HUMAINT project and the Digital Economy Unit (B6) of the Joint Research Center of the European Commission.

COPYRIGHT

CONTACT INFORMATION

David.FERNANDEZ-LLORCA@ec.europa.eu
Emilia.GOMEZ-GUTIERREZ@ec.europa.eu