Užupis Principles for Trustworthy AI Design

"You cannot hope to build a better world without improving the individuals. To that end, each of us must work for our own improvement." - Marie Curie

Like any innovation, Artificial Intelligence (AI) comes with opportunities and threats for humankind. We will only harness AI's wonderful potential, if its benefits increasingly outweigh its harmful uses. Fortunately, many clever brothers and sisters discuss how to orient AI towards common good.

But the ethical debate seems stuck. We either still focus on putting up rules, regulations, and principles to control the people designing AI. Or we try to program our ethical standards into the AI system itself to control the behaviour of AI. Both approaches don't seem to capable of preventing upfront the misuse and unintended consequences of AI. They only make sure that - after something went wrong - there is someone who can be made guilty.

We, the Republic of Užupis, favour trust over control and guilt. We believe that people are fundamentally good. But we also acknowledge that this world is too complex for humans to define ethical standards persistent over space and time.

This is where we would like to start with our "Užupis Principles for Trustworthy AI Design" as a complementary to the European Union's recommendations on trustworthy AI. These are the first principles, which

- rely on trust in the people designing AI,
- support an ongoing process of readjusting the AI, and
- foster adoption to diverse and dynamic conceptions of ethics.

Our principles neither certify nor enforce the design of AI for common good. We rather encourage you to stay true to yourself and act responsibly towards others. These six simple principles shall help you find the right way for you and all the living beings affected by your wonderful work on AI.

- **Trust**: I trust myself that I will use AI to strive for common good whenever possible.
- **Discuss**: I promise to discuss my opinion about what is common good with the various groups affected by my AI.
- **Anticipate**: I promise to do my best to identify anybody and anything affected by my AI and that I will research and anticipate its impact, its uses and its possible misuses.
- **Monitor**: I promise to continuously monitor the impact of my AI and regularly seek feedback from the various groups affected by it.

If necessary

- **Rework**: I promise to revise or even discard my AI if I discover that it did or could cause harm to anybody or anything.
- **Compensate**: I promise to make everybody happy again if my AI made somebody sad for any reason.

If you agree with these principles, we trust you to design AI for common good. Your are welcome to freely use our patch "Trustworthy AI Design" online and offline.

Thank you!