Dual-use of AI-generated content - A question of trustworthiness?

It is undisputedly a goal of today's research and ethics to work towards ensuring that AI-generated content, e.g. texts created by ChatGPT, are fundamentally regarded as trustworthy by users. Whether this goal is achieved depends, of course, on how well the machine learning method works and on the quality of the training data. Nevertheless, it happens that, in addition to the many positive features they offer, AI text generators produce false or distorted information that can contribute to clouding our sense of truth and diminishing our ability to make informed rational decisions. AI text generators therefore have a severe potential to be used for good and for bad. Insofar as this technology presents us with a novel dual-use problem, we can begin to solve it by applying the classical methods of control and prevention explicated in my lecture. However, we must seriously ask ourselves whether an application of these methods of intervention can succeed at all without the appropriate portion of trust. I will try to conclude by outlining the criteria that need to be taken into account in order to launch a trust-based dual-use research on AI.