Preface to the Proceedings of the 1st Workshop on Explainability, Transparency, and Safety (ExTraSafe)

Benjamin Paaßen[1,*]

[1]Bielefeld University, Inspiration 1, 33619 Bielefeld, Germany

1. Introduction

With the increasing deployment of AI systems in application scenarios that profoundly impact humans, demands on the trustworthiness of AI systems also increase, as exemplified prominently in the EU guidelines for trustworthy AI and the EU AI act. As key dimensions of trustworthiness, explainability, transparency, and safety (ExTraSafe) are particularly prominent and extensive AI research has focused on developing novel methods to enhance ExTraSafe properties of AI models and systems. The goal of this workshop was to engage in transdisciplinary discussion between AI research, human-AI interaction, AI practitioners, and law. Further, the workshop functions as annual meeting of the newly introduced work group ExTraSafe of the AI section of the German Computer Science Society (GI - FBKI - AK ExTraSafe).

The workshop took place with the 49th German AI Conference (KI2026) in Bremen, Germany, on August 11, 2026.

2. Call for Papers

The following call for papers was openly distributed across mailing lists of the German Computer Science Society, Connectionists, on social media and via the workshop website.

Key topics of interest are (but are not limited to):

- Legal regulation of AI systems, in particular papers on the legal compliance of AI systems in certain application domains and design implications of legal regulations.
- Practitioners' perspectives on ExTraSafe properties of AI systems in high-stakes application domains, e.g. automotive industry, industrial automation, healthcare, education, or agriculture.
- Analysis of ExTraSafe properties across the entire system lifecycle, especially with respect to architecture considerations, algorithm design, and evaluation practices.
- Approaches to social explainable artificial intelligence (sXAI), taking into account heterogeneous explanatory needs, social roles and contexts, incremental multi-step explanations, and/or multimodality.
- Novel approaches to the evaluation of ExTraSafe properties, especially in user and field studies.

3. Number of Submissions and Peer Review

The workshop received 14 submissions overall, 3 as unpublished discussion entries. Three papers were desk rejected due to lack of fit, the discussion entries were reviewed by at least one reviewer from the list of organizers, all remaining (8) submission were reviewed by two independent reviewers. 6 papers were ultimately accepted for the workshop.

1st Workshop on Explainability, Transparency, and Safety (ExTraSafe)

*Corresponding author.

bpaassen@techfak.uni-bielefeld.de (B. Paaßen)

https://bpaassen.gitlab.io (B. Paaßen)

0000-0002-3899-2450 (B. Paaßen)

4. Workshop Programme

The workshop featured a keynote by Prof. Hannah Ruschemeier (not part of the proceedings) and three sessions with two presented papers, each. The sessions were:

1. Compliance, ethical aspects, and regulation
2. Robustness & Trust
3. Explainability

5. Organizers

The workshop was organized by the AK ExTraSafe of the AI section of the German Computer Science Society (GI - FBKI - AK ExTraSafe). The organization committee was:

- Benjamin Paaßen (Bielefeld University; DFKI), bpaassen@techfak.uni-bielefeld.de (Main Contact Person)
- Vera Schmitt (TU Berlin)
- Thomas Kosch (Humboldt-Universität zu Berlin)
- Julius Schöning (Osnabrück University of Applied Sciences)
- Hanna Drimalla (Bielefeld University)
- Florian Rabe (University Erlangen Nuremberg)
- Tim Schrills (University of Lübeck)
- Daniel Neider (TU Dortmund University; Research Center Trustworthy Data Science and Security; Lamarr Institute for Machine Learning and Artificial Intelligence)
- Peter Fettke (German Research Center for Artificial Intelligence (DFKI); Saarland University)