Baltic Engagement Centre for Combating Information Disorders

REPORT ON ETHICAL ASPECTS OF BECID2

NOVEMBER 2025

MARTEN JUURIK | BECID 2 ETHICS MENTOR

Co-funded by the European Union

SHORT INTRODUCTION

BECID2 is the continuation of the first Baltic Engagement Centre for Combating Information Disorders project co-funded by European Commission. The main aim is to support the work of the engagement centre that was created during the first project. BECID2 shifts its focus on proactive and flexible activities, which is also reflected in the choice of partners: galleries (G), libraries (L), archives (A) and museums (M). The main stakeholder has also been stated to be the general public who is the main target of the communicative, educational and collaborative activities within the project. This also means that BECID2 extends outside of the research ethics framework as many of the activities are not merely research activities (in a strict and narrow sense) but touch upon ethics of other fields like journalism, communication, education and cultural heritage.

The report focuses on three main topics: 1) involvement of humans; 2) use of personal data; 3) use of AI. The report ends with suggestions and general remarks.

The report is based on planned activities of BECID2 – this information comes from official project documentation and materials shared with me. Some of the information may be preliminary and subject to change, since some project activities are yet to begin.

A. INVOLVEMENT OF HUMANS

As a general remark, the topic of involvement of humans can be divided into two broad categories: 1) involvement of human participants in research; 2) involvement of humans in other project activities.

1. RESEARCH PARTICIPANTS

The two clear instances of involving research participants are the focus-group interviews with stakeholders. Informed consent, ethics committee approval, pseudonymisation were mentioned as part of the focus-group design. This looks good, and I don't see any issues raising from these, unless controversial issues are addressed or participant are reminded of their personal negative experiences (e.g. harassment, threats or similar). Based on the materials, however, it seems that most of the participants are in an expert role. However, if non-experts are interviewed about their personal experience with misinformation (especially if health-related) this could raise ethical concerns that are easily mitigated by submitting the research for ethics committee review.

I would point out that the Estonian regulation on research ethics committees is going through a change and from 2026 onward Tartu University's ethics committee would only review research without external funding (at least that's what I've heard). This might be

THIS PROJECT WAS FUNDED BY THE BALTIC ENGAGEMENT CENTRE FOR COMBATING INFORMATION DISORDERS (BECID) 2025–2028;
PART OF MS4 OF THE EU-FUNDED PROJECT WITH THE ID: 101226242, COORDINATED BY THE UNIVERSITY OF TARTU (UTARTU).

something to consider when planning for approval.

2. INVOLVEMENT IN PROJECT ACTIVITIES

Protection scheme against adversary behaviour: this is a crucial and commendable action. I trust that the project team is already well-experienced in this, but nonetheless I would suggest broadening this protection scheme also to the GLAM and educational partners who could lack relevant experience and thus be easy targets of adversary behaviour. There has been a somewhat similar recent incident in Estonia with the head of the Narva Museum. This should be considered when planning activities under T4.2.

In addition, it could be useful to establish a sort of reporting mechanism within the project to document any potential harassment or attacks against the project members or partners during the project. These incidents could be a source of reflection while planning future activities. In the long run, these incidents could be a valuable experience to share with other experts and researchers.

Since the project work involves a lot of university students, these protective measures should also address them and try to prevent any harm befalling them.

Support in abrupt situations: T2.2 covers the aim of providing advice to stakeholders during disinformation campaigns. Since one of the tasks is also to involve vulnerable/marginalised/underserved groups, this raises the issue of addressing the needs and interests of those vulnerable groups during times of crisis (or abrupt situations) when the priority is most likely on dealing with the issue at hand. The project could reflect on the issue of how to deal with abrupt situations without further increasing the vulnerabilities of otherwise marginalised/underserved groups (e.g. how to make sure that they are not last to be informed or "served").

Vulnerable or underserved groups: the mapping of vulnerable groups wasn't done yet by the time of the report. To the best of my understanding, vulnerability is understood mostly as susceptibility to mis-/disinformation, however this trait could coincide with more traditional sources of vulnerability like age, disability, income or marginalisation. Although work with vulnerable individuals or groups require higher ethical standards, the aim of improving participants' MIL capacity likely increases their resilience. Therefore, the risk the vulnerable participants are somehow harmed during the project, is rather low. However, special care should be taken if vulnerable participants are encouraged to take specific actions in social media or in relation to disinformation.

Also, it could be useful to consider whether there are some other language-based minority groups (like local Ukrainian population) that deserve more attention and

THIS PROJECT WAS FUNDED BY THE BALTIC ENGAGEMENT CENTRE FOR COMBATING INFORMATION DISORDERS (BECID) 2025–2028; PART OF MS4 OF THE EU-FUNDED PROJECT WITH THE ID: 101226242, COORDINATED BY THE UNIVERSITY OF TARTU (UTARTU).

involvement within the project. If so, perhaps some outputs could be translated to the relevant languages.

Gendered disinformation: the project documentation rightly points out that this topic is polarising and divisive. It would be good if *counterspeech* and other activities against gendered attacks would be carried out in a gender-balanced way. It should also be considered that opponents of gendered disinformation will very likely become the new targets of the same bad actors. This could reduce the likelihood of finding new spokespersons who are not already well-known and associated with the topic.

It would be great if the project could propose advice or recommendations to individual who want to speak up or share marginal perspectives, but are afraid of online hate or *incivility*.

It would also be great if the project somehow managed to extend the list of spokespersons on gender issues with individuals that transcend (or doesn't fit with) the left-right or conservative-liberal divide. By using the same (easily accessible) spokespersons, there is the risk that existing social divisions will be further entrenched, and the likelihood of bridging social divisions decreases. All in all, it is a very complex issue to address, and I don't have very good recommendations to offer.

Organising public events on controversial topics: this could easily attract attention of disruptive individuals or organisation that could try to undermine the good efforts of the project. There is not much that can be done to prevent this, but the project team should be aware of such a possibility, especially when the events involve young participants.

B. USE OF PERSONAL DATA

In general, I don't have much to say about personal data. Based on the DMP (as of 11.25) there are two main sources of personal data. First, some data are collected directly from participants via interviews. Secondly, some data are collected from social media platforms via APIs or cleanroom type solutions (Meta was mentioned in this regard).

Both collection activities are rather low risk in themselves. In case of interviews, the data subject has greater control over their data by giving and potentially withdrawing consent. In case of social media data, the cleanroom solution usually minimises risks to privacy by limiting access to data for researchers and by allowing only aggregated or generalised data to be exported.

The only point of concern could be use of APIs which could allow collection of different types of identifiable data (as long as the collected data is linkable to specific user accounts). Since the topic of misinformation, especially during elections, is closely related to political views, then at least some of the collected information could be

THIS PROJECT WAS FUNDED BY THE BALTIC ENGAGEMENT CENTRE FOR COMBATING INFORMATION DISORDERS (BECID) 2025–2028;
PART OF MS4 OF THE EU-FUNDED PROJECT WITH THE ID: 101226242, COORDINATED BY THE UNIVERSITY OF TARTU (UTARTU).

considered to belong to special categories under the GDPR art 9. The processing of special categories of personal data without the data subject's consent is permitted for research purposes under certain conditions. However, this could be differently regulated in the different Baltic States.

Risks to privacy: leaving the legal aspects of personal data aside, there is a chance that, incidentally or not, the activities in BECID2 identify certain named individuals who are especially active in disseminating misleading information. This raises the potential issue of whether and to what extent those individuals should be identified to the public or named in any of the BECID materials. Although, shedding light on such instances could have a deterring effect, it is difficult to ascertain whether the individuals knowingly spread disinformation or whether they are victims of disinformation campaigns.

Falling prey to disinformation could be considered shameful, undignified or harmful to reputation, which some people might be reluctant to admit or share freely. Such wishes should be respected whenever it is relevant.

On the other hand, specific illustrative cases might have a potential for educational use (e.g. warn others not to repeat the same mistakes). In those cases, it would be advised to generalise or anonymise the cases so that actual persons wouldn't be revealed. Unless, of course, the disinformation is spread by clearly fake or bot accounts – no protection is warranted then.

OSINT tools rely on personal data: this is just something to keep in mind. When conducting open-source intelligence activities, the sources of information (social media, content platforms) most certainly will contain personal data. Although, this is publicly available, it still falls under the GDPR and requires at least the minimum legal due diligence.

C. USE OF AI

Based on project materials and explanations, BECID2 will use existing tools that are available on the market, and the project would not develop any new AI models or tools. The risks are therefore relatively low as these tools are widely available to anyone.

The specific list of tools was yet not available, however WP1 mentioned use of Copilot and WP3 mentioned use of generative AI for text editing or writing. The risk of generative AI models hallucinating and making things up, is well known and shouldn't be a surprise for the project team. All in all, the general principles of AI ethics should be followed when relevant. If AI is used in the project activities, it should be transparent and explained in clear language. If certain AI tools are verified and suggested to be used by partners or stakeholders (GLAM institutions, schools, educators) the provided materials should also

THIS PROJECT WAS FUNDED BY THE BALTIC ENGAGEMENT CENTRE FOR COMBATING INFORMATION DISORDERS (BECID) 2025–2028; PART OF MS4 OF THE EU-FUNDED PROJECT WITH THE ID: 101226242, COORDINATED BY THE UNIVERSITY OF TARTU (UTARTU).

cover potential risks of those tools.

Risk of misuse: there is a small risk that while introducing different AI tools and manipulations (deepfakes), this could introduce these same tools to youth who would be then tempted to use these tools for the same deceitful or harmful purposes. These tools could be especially harmful when used for bullying. This is just something to consider. The risk could perhaps be mitigated by introducing certain activities to older age groups.

ADDITIONAL REMARKS AND RECOMMENDATIONS

In addition to the previous three topics, I also noticed some further issues that deserve attention. These are minor remarks or ideas that came into mind while I was reading through the descriptions of BECID2 activities.

Communication channels: it has been mentioned (T1.2) that WhatsApp channel is used to announce new outputs. If possible and feasible, alternative channels could be considered as WhatsApp channel requires an account to access and thus limits potential involvement of those who do not want a WhatsApp account. Also, relying to much on this one channel further solidifies its status as a VLOP. WP2 activities directed at the wider public listed additional channels, so most likely this is not an issue. The general recommendation would be to avoid exclusive communication so that certain information is available only in specific channels.

Cooperation with the private sector: T2.4 mentions collaboration with the private sector to use their channels for reaching the public. Considering that Delfi is one of the private media companies, it is not entirely clear with whom collaboration is planned. Since the Baltic media market is quite small, it should be considered how to choose private sector partners in a fair manner.

Also, the project doesn't specifically mention public broadcasters or public owned media. Since, one of the functions of public media organisations is to address or overcome potential "market failures" of private media in reaching certain "underserved" groups. If trying to reach otherwise hard to reach groups, a diversified set of media partners could be considered (as opposed to the most popular ones).

Since BCME is part of the project, perhaps this has already been considered and dealt with, but it wasn't quite clear in the shared documentation.

WP3

Language strategy: T3.1 mentions a language strategy based on which the fact-checks will be translated either into English **or** Russian. The decision will be based on "relevancy

THIS PROJECT WAS FUNDED BY THE BALTIC ENGAGEMENT CENTRE FOR COMBATING INFORMATION DISORDERS (BECID) 2025–2028; PART OF MS4 OF THE EU-FUNDED PROJECT WITH THE ID: 101226242, COORDINATED BY THE UNIVERSITY OF TARTU (UTARTU).

for the specific target group" but it is not clear what does it mean. It is understandable that English translations are more relevant for EDMO network and other European partners or stakeholders, but at the same Russian translations would be crucial considering the local context and the aim of reaching Russian-speaking groups. It is commendable that important information is translated and thus made accessible to a wider public. However, as many fact-checks as possible should be translated to both English and Russian. This would align well with the general goal of the project.

THIS PROJECT WAS FUNDED BY THE BALTIC ENGAGEMENT CENTRE FOR COMBATING INFORMATION DISORDERS (BECID) 2025–2028;
PART OF MS4 OF THE EU-FUNDED PROJECT WITH THE ID: 101226242, COORDINATED BY THE UNIVERSITY OF TARTU (UTARTU).