# A multimodal measurement of the impact of deepfakes on the ethical reasoning and affective reactions of students

**V Ramachandran**[1]
The Teaching Support Center (CAPE), Ecole Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
https://orcid.org/0000-0001-5249-2578

**C Hardebolle**
Center for Digital Education (CEDE), Ecole Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
https://orcid.org/0000-0001-9933-1413

**N Kotluk**
Center for Learning Sciences (LEARN), Ecole Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
https://orcid.org/0000-0002-4314-9492

**T Ebrahimi**
Multimedia Signal Processing Group, Ecole Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
https://orcid.org/0000-0002-9900-3687

**R Riedl**
Institute for Digital Technology Management, Berner Fachhochschule
Lausanne, Switzerland
https://orcid.org/0000-0002-4483-9997

**P Jermann**
Center for Digital Education (CEDE), Ecole Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
https://orcid.org/0000-0001-9199-2831

**R Tormey**
The Teaching Support Center (CAPE), Ecole Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
https://orcid.org/0000-0003-2502-9451

**Conference Key Areas**: Embedding Sustainability and Ethics in the Curriculum, Education about and education with Artificial Intelligence

**Keywords**: Artificial intelligence, Deepfakes, Emotions, Ethics, Moral judgement

---

[1] *Corresponding Author*

*V Ramachandran*

*vivek.ramachandran@epfl.ch*

**Abstract**

Deepfakes - synthetic videos generated by machine learning models - are becoming increasingly sophisticated. While they have several positive use cases, their potential for harm is also high. Deepfake production involves input from multiple engineers, making it challenging to assign individual responsibility for their creation. The separation between engineers and consumers may also contribute to a lack of empathy on the part of the former towards the latter. At present, engineering ethics education appears inadequate to address these issues. Indeed, the ethics of artificial intelligence is often taught as a stand-alone course or a separate module at the end of a course. This approach does not afford time for students to critically engage with the technology and consider its possible harmful effects on users. Thus, this experimental study aims to investigate the effects of the use of deepfakes on engineering students' moral sensitivity and reasoning. First, students are instructed about how to evaluate the technical proficiency of deepfakes and about the ethical issues associated with them. Then, they watch three videos: an authentic video and two deepfake videos featuring the same person. While watching these videos, the data related to their attentional (eye tracking) and emotional (self-reports, facial emotion recognition) engagement is collected. Finally, they are interviewed using a protocol modelled on Kohlberg's 'Moral Judgement Interview'. The findings can have significant implications for how technology-specific ethics can be taught to engineers, while providing them space to engage and empathise with potential stakeholders as part of their decision-making process.

## 1 Background

In this paper, we introduce a mixed-methods measurement method that allows us to study the effects of educating engineering students about the ethical and technical aspects of a specific technology on their moral judgement. The technology here is deepfakes, a form of Generative Artificial Intelligence (AI), which are synthetic videos created using machine learning (ML) models. As part of this study, we create a space for students to articulate their emotional experience and for us to capture their attentional data when watching deepfake videos. We believe that this approach may explain how students apply their ethical education when engaging with stakeholders as an important step towards developing their moral judgement.

### 1.1 Rise of Deepfakes - the promise and the danger

The fields of ML and AI have made tremendous advances over the past decade, particularly in computer vision, computational linguistics, and human-computer interaction (Wang and Keng, 2019). These advances have been made possible due to a combination of novel, sophisticated ML algorithms, large multimedia datasets, and powerful graphics-related hardware to optimise training.

Generative AI refers to a class of AI predictive methods that can generate different types of data in the form of synthetic media - text, image, audio, video - using existing data of the same format (Westerlund, 2019). Deepfakes are a type of synthetic media, often audio/video, produced by a combination of generative AI. Deepfake is a portmanteau of "deep learning", a class of ML algorithms involving the use of artificial neural networks and "fake", as in unreal. Deepfake media are created by manipulating or replacing the original audio/video with fabricated or altered content, often making it difficult to discern the authenticity of the resulting media (Karnouskos, 2020). Broadly speaking, deepfakes can be categorized into three types:

1. Head puppetry/Face swapping – A video is created to show a synthesized person's head and shoulders that mimics the behaviour of a real person's head movements. Video of the real person is used as source material for deepfake creation. In some instances, the synthesis person can be used on another real person's face.
2. Lip-syncing – A video is created with new audio by manipulating the lip movement of the person's face in the source video such that in the final deepfake video, the person appears to say something different to what they said in the original video.
3. Voice Cloning – This technique is used to generate audio-only media, in which a simulated voice is created based on multiple audio samples of the real person's voice such that the simulated voice is similar in sound to the real person.

Deepfakes have been used for numerous creative and constructive purposes in a wide range of avenues including healthcare, commerce, fashion, and education (Westerlund, 2019). At the same time, they are more commonly associated with producing content that ranges from hilarious to nefarious. Indeed, there are significant ethical issues to be grappled with in addressing the threats they pose to the public (Whittaker, 2020). Specifically, deepfakes have been used to misrepresent

individuals and misinform the public. Being a target of a deepfake can also lead to loss of trust and credibility, as false actions or statements attributed to individuals can spread rapidly through social media, creating confusion and misinformation. The potential harm to individuals' personal and professional lives as a result of being targeted by deepfakes is gravely concerning. Indeed, multiple commercial as well as free software are easily accessible that allow users to create life-like deepfakes regardless of their intended purpose. The number of successful companies developing this kind of technology is increasing. They are recruiting skilled engineers, who could contribute to the incorporation of ethical thought in their practice. Therefore, there is an urgent need to train engineers about the ethical issues with deepfake technology, which we detail in the next section.

### 1.2 Lack of Responsibility

The production and use of deepfakes can involve a number of steps, with inputs from a wide variety of actors. This includes the engineers who develop the algorithm, create the datasets, train the model, test it for specific applications, before releasing it to the public who can then customise the technology to create deepfakes for the applications of their choice. Therefore, when individuals are targets of misrepresentation due to non-consensual creations of deepfakes, the responsibility of this harm becomes difficult to attribute to one person alone - this is a classic example of the “many hands” problem in engineering ethics in which the attribution of individual responsibility becomes extremely difficult in collective settings (Van de Poel and Roayakkers, 2011).

In addition to the difficulty of determining accountability, there is an added effect of the distance between the engineers who develop the algorithm and the one who is “deepfaked”. This separation between the producers of the deepfake technology and those affected by their production increases the risk of producers feeling released from the traditional social obligations towards the latter. There is a perception commonly held among some engineers involved in technology development that the responsibility for the consequences of the technology's use falls on others rather than themselves (Isaac et al., 2023).

Moreover, prevailing ideas of software technologies being objective (or a net positive) and unaffected by the values of the developer allow producers of deepfakes to free themselves of social obligations to those affected by deepfake dissemination (Griffin et al., 2023). This is compounded by the prevailing notion that ethics is a management issue and not an engineering one.

While computer engineers may consciously or subconsciously not consider their ethical responsibility, the technologies that they create have major ramifications in terms of the negative effects on the individuals who are being deepfaked. It can result in reputational damage, emotional distress, and violation of privacy, particularly in cases of revenge porn where someone's face can be swapped onto explicit content without their consent.

### 1.3 Ethics Education in Computer Engineering

In recent years, there has been a spate of novel approaches to integrate ethics in software and computer engineering curricula that would seek to introduce ethics at multiple levels of the study program (Horton et al., 2022; Grosz et al., 2019). However, in most computer engineering programs, ethics is taught as a stand-alone course as part of the department’s sole ethics course (Fiesler et al., 2020) In some other programs within computer science, ethics is taught in different courses but as a separate module (Grosz et al., 2019). Moreover, ethics education is often interchangeable with teaching the Code of Ethics as prescribed by different professional organizations in the discipline (Fiesler et al., 2020). While these are all important and necessary efforts towards creating more ethically-minded engineers, they are not sufficient because they do not provide enough time for students to critically engage with the ethical aspects of each technology concept that is taught to them. In order to facilitate this critical engagement, the ethical issues of specific technologies need to be taught along with the technology itself (Martin et al., 1996). This form of intervention would allow educators to emphasize the social responsibility of engineers as technology creators throughout the curriculum.

Indeed, there is a growing need for this form of intervention to address the ethical concerns associated with deepfakes in engineering education. As deepfake technology becomes more advanced and accessible, engineering students and professionals are increasingly able to create realistic fake media (Kietzmann et al., 2020). This raises ethical concerns related to the potential misuse of deepfakes, such as spreading misinformation, manipulating data, and violating privacy and consent. Engineering

education programs must incorporate discussions on the ethical implications of deepfakes, including the responsible use of the technology and the potential consequences of its misuse.

Therefore, it is important to intervene as early as possible in engineering education to instil this sense of responsibility amongst engineering students towards potential stakeholders of their technological creation, such as deepfake technology, so that they can develop the ability to make ethical decisions. An important consideration in fostering ethical decision-making skills is that it is an entirely cognitive exercise i.e., our ethical decisions are defined by a combination of factors, including emotional relationships with oneself and others (Riley, 2013).

## 2 Experimental Design

In this study, we aim to compare the effects of a computing education topic, here deepfakes, that includes both technical and ethical aspects with one with a purely technical education on engineering students. Specifically, we are interested in two assessing two effects. One, the impact of the educational content on - 1) students' attention and emotionality with respect to their engagement with authentic and deepfake representations of a person and, 2) students' moral judgement in ethically ambiguous situations. Our proposed method consists of a human subject study with three phases - an *Education phase*, an *Engagement phase,* and an *Interview phase*.

### 2.1 Education Phase

The purpose of the *Education Phase* is to provide a short education to subjects regarding specific aspects of deepfake technology. The technical aspects put emphasis on learning what deepfakes are, how to recognize them, identify common audiovisual artefacts, and distinguish between genuine and manipulated media. Deepfakes present a multitude of ethical dimensions, but our intent is to highlight one in particular - the relationship between the technology creator (the subject) and the target/unintended stakeholder (the person who is deepfaked). Thus, the ethical aspects of the educational content are centred on the profound impact of deepfakes on targeted individuals through their non-consensual misrepresentation. The ethical education is created with the intention of fostering empathy and instilling a sense of responsibility in the subject towards the victims of deepfake manipulation. At the end of the *Education Phase*, the subject should have a clear idea of their role as technology creators, which should enable them to make informed and responsible decisions.

### 2.2 Engagement Phase

The *Engagement Phase* is designed to give subjects an opportunity to connect with an individual, targeted by deepfake creation. The subjects watch an authentic/unaltered video of an HR person giving a recruitment talk for their engineering company. The unaltered video allows subjects to witness the HR person's genuine form and expressions in an unmanipulated context so that subjects develop a baseline understanding of their appearance and demeanour. Then, the subjects watch two deepfake versions of the authentic video in which both the audio and video have been altered, whereby the altered recruitment talks differ from the unaltered one in terms of the person's speech, tonality, and sincerity. The purpose of making the subjects watch two different types of deepfake videos is threefold.

One, subjects are asked to assess the quality and content of the deepfake videos they observe. The subjects' evaluations help us understand how they evaluate the authenticity of the manipulated media. The subjects' attention to common deepfake-related production flaws is also an important indicator because these artefacts may raise suspicion towards the person in the video and/or be distracting from what the person says.

Two, the videos let the subjects reflect on their social emotional response toward the person in the video based on a three-axis model developed to ascertain student-relationships in classrooms i.e. assertion, affiliation, and attachment (Tormey, 2021). Therefore, this measurement enables the subjects to express their relationship towards the person in each of the videos. Since the video is a recruitment talk, this indicator may describe the subjects' perceptions of the HR person's professionalism.

Three, the videos also allow the subjects to introspect their culpability as potential creators of the video i.e., when confronted with the possibility that they had a role in creating these deepfake videos, the

subjects' have an opportunity to express their moral emotions that they experience (Haidt, 2013). These moral emotions are a combination of emotions that are self-conscious emotions and those projected on the deepfaked person.

In summary, the *Engagement Phase* allows subjects to apply their knowledge from the *Education Phase* to distinguish between genuine and manipulated media, express their emotionality and attention towards the person in each video, and confront the potential ethical implications and consequences of their actions as deepfake producers.

### 2.3 Interview Phase

As part of the *Interview Phase*, our primary objective is to assess the subjects' level of moral judgement by employing Lawrence Kohlberg's framework of Moral Development that comprises three stages: Pre-conventional, Conventional, and Post-conventional (Kohlberg, 1971). To investigate the impact of ethics education on their moral development, we are interested in understanding the effects of undergoing the *Education* and *Engagement Phases*. We can gauge the potential influence of the educational interventions on their ethical decision-making abilities and the evolution of their moral outlook by evaluating their moral reasoning and judgement throughout the study.

We use the Neo-Kohlbergian Defining Issues Test, specifically adapted to Engineering Sciences, known as the Engineering and Science Issues Test (ESIT) for this examination (Borenstein et. al., 2010; Kotluk and Tormey, 2022). This test involves presenting the subjects with a set of three distinct cases, each highlighting an ethical dilemma. These cases revolve around the creation, utilization, and dissemination of deepfake technology as well as its potential impact on individuals. During the test, the subjects are required to read and analyse each case individually. Subsequently, they are prompted to make moral judgements and offer justifications for their decisions. Specifically, we try to ascertain their sensitivity to the issue presented in the case, their motivation to address it urgently, and their reasoning in identifying decision criteria.

By employing the ESIT, we can observe the subjects' ethical considerations that guide their decision-making and thus evaluate the subjects' moral development. This tailored assessment tool enables us to assess their moral judgements, reasoning abilities, and the ethical frameworks they employ when confronted with complex dilemmas involving the creation and use of deepfake technology. Through this comprehensive analysis in the *Interview Phase*, we aim to shed light on the potential impact of ethics education on the moral sensitivity and reasoning of engineering students

## 3 Methodology - Implementation and Data Collection

In this section, we describe how the three different phases are implemented and what methods are used to collect data during each phase. As mentioned earlier, we conduct a human subject study to investigate the effect of teaching ethical aspects of a computing education topic alongside technical aspects on the moral development of subjects. The computing education topic is deepfakes - its production, detection, and impact on targeted individuals. We pay specific attention to the role of engaging with deepfaked individuals on subject attention and emotionality.

To ensure a controlled environment, we recruit engineering students (bachelor's and master's), who possess a limited understanding of deepfakes. The decision to select subjects based on this criterion is motivated by our desire to assess the impact of our interventions on individuals who may not be fully familiar with the intricacies of this emerging technology, but are most likely to be potential creators of similar AI technologies. The subjects participating in the study are divided into two groups: the control group and the test group. As shown in **Figure 1,** the subject study is conducted in the three phases in chronological order, one subject at a time. During the *Education Phase*, the control group receives only the technical education, whereas the test group receives both the technical and ethical education about deepfakes. During the *Engagement* and *Interview Phases*, both groups receive identical treatment, ensuring that the observed differences can be attributed to the intervention in the *Education Phase*.

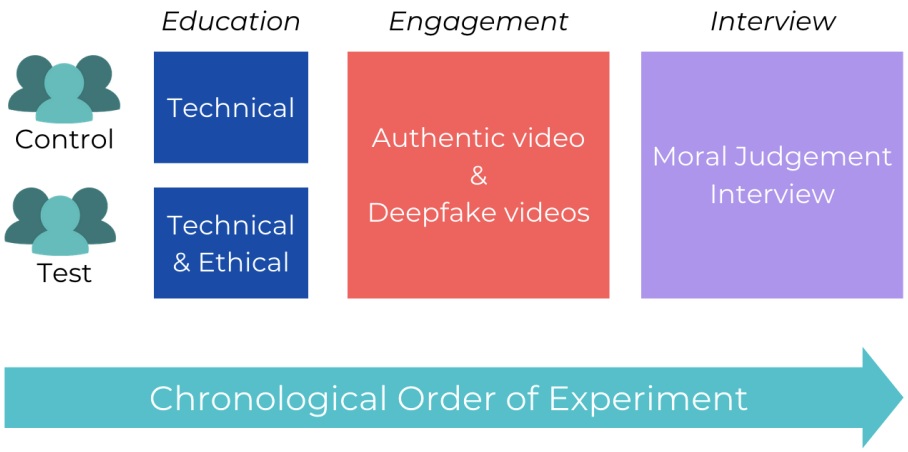


**Figure 1**: The chronological order of the subject study for each subject consists of three phases - Education, Engagement, and Interview

During the experiment, the subject is seated at a table facing a monitor with access to a mouse and a keyboard. In addition, an eye tracker (Tobii Pro Fusion) is mounted on the monitor and calibrated for each subject at the beginning of the experiment. Also, a video camera is mounted on a separate stand behind the monitor and positioned such that it records each subject's face. A separate monitor, keyboard, and mouse is placed adjacent to the subject's monitor, which is controlled by the experimenter. **Figure 2** shows the experimental setup that is used for running the subject study. The human subject study is approved by the EPFL Human Research Ethics Commission (HREC), provided that subjects' give their informed consent to participating in the study.

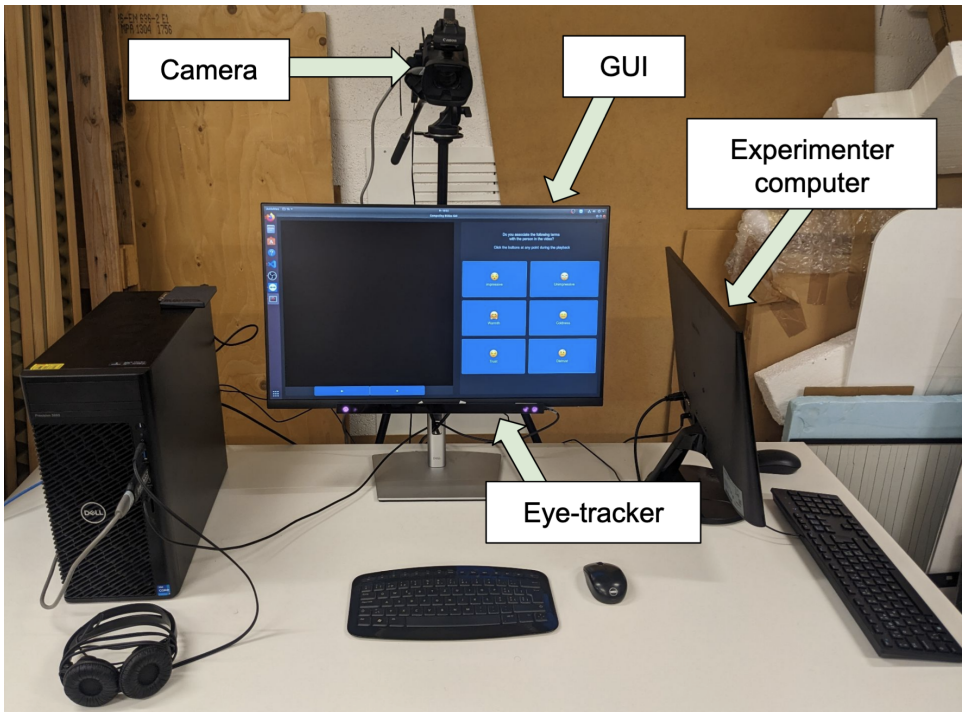


**Figure 2**: The experimental setup used to perform the human subject study, consisting of a computer for the subjects to interact with a graphical user interface (GUI), a computer for the experimenter, an eye-tracker to monitor their eye gaze movement, and a camera to record their facial expressions.

### 3.2 Education Phase

In the *Education Phase*, the control group exclusively receives technical education focused on what deepfakes are and how they can be detected. Subjects watch a video, entitled Education Video 1, nested in a graphical user interface (GUI) featuring the experimenter. During the video, the experimenter describes what defines deepfakes are and how they are created. Furthermore, the experimenter provides key steps to distinguish manipulated media like deepfakes from genuine media by paying attention to audiovisual artefacts. The script used for Education Video 1 is as follows:

"*Deepfakes are synthetic media created using deep learning algorithms to replace or superimpose a person's image, voice, or behaviour with that of another. Here are the steps to detect a deepfake. Firstly, look for any visible distortions or glitches in the video, such as mismatched lighting or blurry edges around the face. This can be a tell-tale sign that the video has been digitally manipulated. Secondly, pay attention to the audio quality. Deepfake algorithms often struggle to replicate natural speech patterns, resulting in distorted or robotic-sounding speech. Thirdly, analyse the movements of the subject in the video. Are the movements smooth and natural or do they appear jerky or robotic? If the movements seem stiff or unnatural, it could be a sign that the video is a deepfake.*"

The test group subjects receive a technical and an ethical education about deepfakes, in that order. The subjects in this group also watch Education Video 1, followed by a second video, entitled Education Video 2, in which the experimenter describes the real-world effects of deepfake technology on the lives of individuals who are misrepresented, often without their consent. The purpose of Education Video 2 is to instil a sense of responsibility and ethical awareness in combating the detrimental effects of deepfakes. The script for Education Video 2 is as follows:

"*Deepfakes are not just a technological issue, they also raise significant ethical concerns, particularly in terms of how they can affect the person who is deepfaked. While deepfakes can be used for harmless entertainment purposes, they can also be used to harm individuals in various ways. A deepfake video can be used to defame or embarrass a person, by depicting them engaging in illegal or unethical behaviour. This can have serious consequences for the person's reputation and may impact their personal and professional life. Even if the deepfake is proven to be fake, the damage may already have been done. Furthermore, the process of creating deepfakes often involves using images or videos of real people without their consent, which raises privacy concerns. This can be particularly distressing for individuals who have been victims of revenge porn or other forms of online harassment.*"

### 3.3 Engagement Phase

The *Engagement Phase* is meant ton assess the effectiveness of imparting the information in the *Education Phase* on subjects' ability to detect deepfakes and recognize ethical issues through gaze-based attention and emotional expression. This phase involves subjects watching three videos - Engagement Video 1, Engagement Video 2, and Engagement Video 3. These videos feature the same individual, an HR person from an engineering company called Protos. The subject watch each of these videos in chronological order. Engagement Video 1 is authentic, and it features an HR person promotes the company's open positions. Measurements pertaining to this video provides a baseline for subjects' attention and emotional response. The script for Engagement Video 1 is as follows:

*"I am an HR Manager at Protos. We are an engineering company that is looking for talented individuals to join our team, and today I am here to tell you about the exciting opportunities we have available. At Protos, we value innovation, creativity, and hard work. We believe that by hiring the best people, we can achieve great things together. That's why we are looking for individuals who are passionate about engineering and who want to make a difference in the world. We offer a range of positions across various departments, from software development to mechanical engineering. Whether you are a recent graduate or an experienced professional, we have something for you. At Protos, we are committed to creating a diverse and inclusive workplace where everyone can thrive. We believe that diversity brings fresh perspectives and new ideas, and we are committed to ensuring that everyone feels welcome and valued. So, if you are looking for an exciting career in engineering, I encourage you to apply for one of our open positions. Thank you for considering Protos as your potential employer. We look forward to hearing from you soon."*

In comparison, Engagement Videos 2 and 3 are deepfakes that misrepresent the HR person in different ways. Engagement Video 2 is meant to portray the HR person as disingenuous, whereas Engagement Video 3 is meant to depict them as incompetent. These different scenarios allow us to compare the measurements with the baseline for Engagement Video 1. In each of the deepfakes, visible production flaws are present in the form of artefacts around the person's mouth, facial

expressions, and tonality to clearly indicate to the subject that they are watching deepfakes. The scripts for Engagement Video 2 and 3 are as follows:

*"I am an HR Manager at Protos. We are an engineering company that is looking for talented individuals to join our team, and today I am here to tell you about the exciting opportunities we have available. At Protos, we say that we value innovation, creativity, and hard work. The company website says that by hiring the best people, we can achieve great things together. That message is important for the public. That's why we are looking for individuals who are going to help us create that impression for our customers. We offer a range of positions across various departments, from software development to mechanical engineering. Whether you are a recent graduate or an experienced professional, we have something for you. At Protos, we say that we are committed to creating a diverse and inclusive workplace where everyone can thrive. That is what all companies (are supposed to) say these days … that we believe that diversity brings fresh perspectives and new ideas. Once again, this will be an important message for our employees to show to the public. So, if you are looking for a career in engineering, and can see the importance of creating the right impression for customers, I encourage you to apply for one of our open positions. Thank you for considering Protos as your potential employer. We look forward to hearing from you soon."*

*"I am an HR Manager at Protos, We are "um" … like an engineering company? that is looking for A "um" … like talented individuals to join our team? At Protos, we value, like…, innovation, creativity, and hard work,…that kind of thing. We believe that by hiring the best people, we can … kind of … achieve ,like … uhhh… great things together. That's why we are looking for individuals who are …umm… passionate about … uhhh … engineering and who want to make a difference in the world. Hmm …. we offer a range of positions across various departments, from, like, software development to mechanical engineering. Whether you are a recent graduate or, like, an experienced professional, we have something for you …hmmm. At Protos, we are committed to creating …ummm … like, a diverse and, uhh ... sort of, inclusive workplace where everyone can thrive. We believe that, like, diversity brings fresh perspectives, kind of, and new ideas. So, if you are looking for an exciting career in engineering, I encourage you to apply for one of our open positions? Thank you for considering Protos as your potential employer. We look forward to hearing from you soon."*

For each video, two types of measurements are made - concurrent (during the video) and terminal (end of the video). Concurrent measurements provide real-time tracking of subject attention and emotional reactions towards the person in the video. Terminal measurements provide aggregate responses from subjects about their self-conscious and outwardly projected emotions, as well as their technical evaluation of deepfake quality and content. While concurrent measurements capture initial, unadulterated perceptions, whereas terminal measurements include refined responses subject to reflection.

### 2.3.1 Concurrent Measurements

There are three concurrent measurements made during each of the Engagement Videos:

1. The subjects express their social emotions towards the HR person in the video by clicking labelled emoticons in response to the HR person's actions/statements in real-time. These emoticons are categorized into three axes: attachment (trust/distrust), affiliation (warmth/coldness), and assertion (impressive/unimpressive) (Tormey, 2021). By using these emoticons, we can gather immediate insights into the subjects' emotional reactions and perceptions of the HR person's credibility, likeability, and competence.
2. The subjects faces are recorded using the camera, which are then processed by an open-source, locally installed facial expression recognition software, called EmoInfer (Sinha and Dhandhania, 2022). EmoInfer analyses and classifies each frame of the recorded video into stereotypical facial expressions as defined by existing facial emotion models. This data may give us some insight into subjects' level of engagement and possibly their subconscious emotional responses.
3. The subjects' eye gaze movement is tracked using the eye-tracker to ascertain which specific aspects of the video they pay attention. This data allows us to identify Areas of Interest and eye gaze fixations, which can indicate their ability to detect deepfake artefacts, as well as to convey emotional empathy to the person in the video.

### 2.3.2 Terminal Measurements

There are four terminal measurements that are made after each of the Engagement Videos:

1. Similar to the concurrent measurement of social emotions, the subjects respond to a questionnaire that uses the same three-axes model. For each axis, subjects rate the person's

characteristics on a 7-point Likert scale, from “Not at all” to “Very much” - trustworthiness, well-intentioned, reassuring, reliable, inspires confidence (attachment); friendly, warm, compassionate, positive towards viewer, caring (affiliation); and impressive, admirable, influential, exciting, inspiring (assertion). This measurement helps us to capture their subjective evaluation of the HR person's social attributes and emotional impact.

2. To complement the quantitive facial expressions recognized by EmoInfer, subjects watch the video recording of their faces alongside a time-synched screen recording of the GUI with the Engagement Video they watched. While watching the recordings, the subjects free-label their own facial expressions as they might be able to better recognize them.
3. For Engagement Videos 2 and 3 , subjects complete a questionnaire that tests their technical proficiency in engaging with a deepfake. They evaluate the quality of the videos - both visual (video) and auditory (audio) aspects, using a 7-point Likert scale. They also indicate the extent to which they paid attention to the quality and content of the video. Finally, they state whether they are able to detect that the video is a deepfake. This self-assessment helps us understand their confidence level in their deepfake detection skills.
4. For Engagement Videos 2 and 3, subjects respond to a questionnaire that pre-supposes their involvement in the creation of the deepfake videos. Based on this supposition, the subjects express their moral emotions - guilt, shame, embarrassment, pride, compassion, contempt, and disgust - on 7-point Likert scales. Their responses provide insight into their ethical sensitivity when confronted with the potential scenario of creators of harmful technology.

This comprehensive assessment comprising concurrent and terminal measurements enhances our understanding of the subjects’ experiences and allows us to draw meaningful conclusions about the effectiveness of the educational interventions and the impact of deepfakes on individuals.

### 3.4 Interview Phase

In the *Interview Phase*, the subjects read three ESIT-type cases, one at a time, that present ethical dilemmas related to the production, usage, and dissemination of deepfakes targeting individuals. After reading the case, the subjects must answer two questions to measure moral sensitivity and motivation:

1. Moral Sensitivity - “Is there an ethical issue in the case you just read? If you respond yes, then what is the ethical issue?”
2. Moral Motivation - “If you have identified an ethical issue, is there an urgency in addressing this issue? If you respond yes, then please elaborate.”

To measure their moral reasoning, subjects are presented with a set of 12 questions that relate to different levels of moral judgement. Using a Think Aloud Protocol (Bernadini, 2001), they are asked to evaluate the relevance of each question to the ethical dilemmas presented in the cases. Subjects select one of five options (“great”, “much”, “some”, “little”, “no”) to indicate the relevance of each question. This allows the subjects to critically engage with each criterion. Finally, subjects select four of the most important questions, in order, that are relevant to the case they read.

Their selections are used to calculate a numeric measure of post-conventional moral reasoning for each subject based on the ESIT scoring key. This scoring key assigns values to different levels of moral reasoning, allowing for a quantitative assessment of subjects' ethical decision-making processes. This quantitative measure is complemented by a qualitative analysis of the subjects' verbal responses that are recorded. Collectively, we can gain valuable insights into their moral sensitivity, moral motivation, and moral reasoning in evaluating ethically ambiguous situations involving the targeted deepfaking of individuals.

## 4 Conclusion

This paper describes a methodology that aims to investigate the effect of ethical and technical education of deepfakes on subjects' deepfake detection skills, their attention and emotionality towards deepfaked individuals, and their moral judgement in ethically ambiguous cases. Presently, we are in the process of collecting data for the proposed study. While we have not made explicit hypothesis in this paper, we posit that the ethical education will promote test group subjects' moral sensitivity, motivation, and post-conventional reasoning. Furthermore, it may highlight test group subjects' ethical tendencies and encourage them to have more empathy towards deepfaked individuals than control group subjects through their gaze and emotional expression.

However, it is important to recognize that the results from this study may be difficult to generalize for a few reasons. One, the length of the experiment is approximately 60 minutes per subject and the *Educational Phase* lasts 15 minutes. Any transferable effects observed in this short timeframe to a classroom setting will need to be verified in a separate longitudinal study. Two, the social identity of the experimenter and that of the HR person may have unforeseeable effects on the results, depending on the social identity of each individual subject. Three, facial expression recognition, machine-read or self-reported, is a heavily contested measure because it is incumbent upon accepting the premise that there are specific universal emotions. While this is not an exhaustive list of possible limitations of this study, we believe that incorporating diverse data collection methods should help offset some of the challenges they pose.

We anticipate that a comprehensive mixed-methods data analysis will contribute to our understanding of the issues posed by deepfakes and other types of generative AI. Through this study, our aim is to develop novel and responsible uses of AI tools in education, especially to teach ethics to engineering students. Ultimately, the insights gained from this study should inform future educational initiatives and empower individuals to navigate the complex landscape of digital media with greater resilience and discernment.

## 5 Acknowledgements

We would like to thank our colleagues in CAPE, CEDE, and LEARN at EPFL for their valuable comments and contributions that have enriched the quality of this paper, in particular the immense support of Jessica Dehler Zuffrey, Iris Capdevila, Matthew Goodman, Magali Croci, Christian Vonarburg, Alexandra Niculescu, and Siara Isaac. This project is funded by the Canton of Bern through the BeLEARN hub.

## References

Bernardini, Silvia. "Think-aloud protocols in translation research: Achievements, limits, future prospects." *Target. International Journal of Translation Studies* 13, no. 2 (2001): 241-263.

Borenstein, Jason, Matthew J. Drake, Robert Kirkman, and Julie L. Swann. "The engineering and science issues test (ESIT): A discipline-specific approach to assessing moral judgement." *Science and Engineering Ethics* 16 (2010): 387-407.

Fiesler, Casey, Natalie Garrett, and Nathan Beard. "What do we teach when we teach tech ethics? A syllabi analysis." In *Proceedings of the 51st ACM technical symposium on computer science education*, pp. 289-295. 2020.

Griffin, Tricia A., Brian Patrick Green, and Jos VM Welie. "The ethical agency of AI developers." *AI and Ethics* (2023): 1-10.

Grosz, Barbara J., David Gray Grant, Kate Vredenburgh, Jeff Behrends, Lily Hu, Alison Simmons, and Jim Waldo. "Embedded EthiCS: integrating ethics across CS education." *Communications of the ACM* 62, no. 8 (2019): 54-61.

Haidt, Jonathan. "The moral emotions." *Handbook of affective sciences* 11, no. 2003 (2003): 852-870.

Horton, Diane, Sheila A. McIlraith, Nina Wang, Maryam Majedi, Emma McClure, and Benjamin Wald. "Embedding Ethics in Computer Science Courses: Does it Work?." In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education V. 1*, pp. 481-487. 2022.

Isaac, Siara Ruth, Aditi Kothiyal, Pierluca Borsò-Tan, and Bryan Alexander Ford. "Sustainability and Ethicality are Peripheral to Students' Software Design." *International Journal of Engineering Education* 39, no. 3 (2023): 542-556.

Karnouskos, Stamatis. "Artificial intelligence in digital media: The era of deepfakes." *IEEE Transactions on Technology and Society* 1, no. 3 (2020): 138-147.

Kietzmann, Jan, Linda W. Lee, Ian P. McCarthy, and Tim C. Kietzmann. "Deepfakes: Trick or treat?." *Business Horizons* 63, no. 2 (2020): 135-146.

Kohlberg, Lawrence. "Stages of moral development." *Moral education* 1, no. 51 (1971): 23-92.

Kotluk, Nihat, and Roland Tormey. "Emotional empathy and engineering students' moral reasoning." *Towards a new future in engineering education, new scenarios that European alliances of tech universities open up* (2022): 458-467.

Martin, C. Dianne, Chuck Huff, Donald Gotterbarn, and Keith Miller. "Implementing a tenth strand in the CS curriculum." *Communications of the ACM* 39, no. 12 (1996): 75-84.

Riley, Donna. "Hidden in plain view: Feminists doing engineering ethics, engineers doing feminist ethics." *Science and engineering ethics* 19 (2013): 189-206.

Sinha, Tanmay, and Dhandhania, Sunidhi. "Democratizing Emotion Research in Learning Sciences." In *International Conference on Artificial Intelligence in Education*, pp. 156-162. Cham: Springer International Publishing, 2022.

Tormey, Roland. "Rethinking student-teacher relationships in higher education: a multidimensional approach." *Higher Education* 82, no. 5 (2021): 993-1011.

Van de Poel, Ibo, and Lambèr Royakkers. *Ethics, technology, and engineering: An introduction*. John Wiley & Sons, 2023.

Wang, Weiyu, and Keng Siau. "Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity: A review and research agenda." *Journal of Database Management (JDM)* 30, no. 1 (2019): 61-79.

Westerlund, Mika. "The emergence of deepfake technology: A review." *Technology innovation management review* 9, no. 11 (2019).

Whittaker, Lucas, Tim C. Kietzmann, Jan Kietzmann, and Amir Dabirian. "'"All around me are synthetic faces": the mad world of AI-generated media." *IT Professional* 22, no. 5 (2020): 90-99.