moonshot

Ethics Framework

JANUARY **2023**

MOONSHOTTEAM.COM

↘ Introduction

This document has been developed by Moonshot's Ethics Committee as one of the foundational documents for program design and delivery. It is updated periodically to reflect methodological and programmatic developments. While many of the questions and considerations here are very specific to our work on online public safety, we hope it will be helpful to others - either in whole or in part. If you find it useful and/or want to share feedback, please do get in touch with us at **Moonshot Ethics**.

The role of this framework

To provide guidance and support to Moonshot staff in designing and implementing ethically sound and responsible work

To provide a systematic, traceable approach to ethical reviews of clients, methodologies, and unintended impacts

To serve as an early warning system for uncovering ethical vulnerabilities, and support proactive solution-oriented thinking

To foster a culture of challenge within Moonshot in line with its goal to serve as a thought leader in online safety

Your role

Project managers are responsible for completing and updating the Ethical Process Framework - and seeking help where needed

This is not a project management or data privacy framework; those requirements are addressed in other internal documents

Ethical design requires you to exercise your judgement and discretion. We trust you to do that - and if you're not sure about the answer, all you need to do is ask.

Ethical questions and considerations for all Moonshot projects

01
Necessity and proportionality

Explain why the work is *necessary* to help safeguard the online space, and how our solution is *proportional* to the public safety threat in question.

> *E.g. is there a risk of threat amplification if the work was to be carried out? Could we amplify one threat at the expense of another, more pressing one, leading to stigmatisation of particular communities? Are we gathering more data than we need?*

02
Social impact and Do No Harm

A. Could project outputs (data, capacity building, local partner financing) be used in a way that creates harm?

> *E.g., could OSINT training be redeployed to surveil individuals? Could the project's partnerships inadvertently inflame existing tensions in the local context?*

B. How have you defined the target audience? What are your risk and targeting criteria, and how are you ensuring these are consistently applied?

> *To consider: Is there a risk of stigmatising a specific ethnic, social, or political group? Are your targeting criteria sufficiently narrow? Are you inadvertently applying geo-demographic criteria (e.g. by focusing on a specific region)?*

C. What is your communications plan? Where are the project outputs being used?

D. Do you have a sustainability or exit plan in place?

> *To consider: risks of exit to local partners and vulnerable individuals engaged by the project; ensure a plan is in place to mitigate those risks.*

03
Human rights

How will the project ensure respect for users’ rights to freedom of expression, belief, and privacy?

04
Inclusivity and diversity

How are you ensuring diversity of perspectives in project design? What internal and/or external peer review/QA processes do you have in place? Can you identify any indicators of unconscious bias influencing the activities and delivery of the project? How are you planning to mitigate this risk - e.g. through internal/external reviews, including by local experts?

05
Data integrity (see also: protection of individual right to privacy under 3)

A. Have you reviewed the **Secure Research Protocol**?

B. Have you completed a **Data Privacy Impact Assessment** and **General Data Protection Regulation (GDPR) Assessment**?

C. For new areas of work, have you actively checked whether there are **additional sector and/or international ethical data collection standards**?

06
Duty of care

A. Have proposed partners/consultants/beneficiaries been vetted based on ethical principles?

Existing Moonshot systems can support ethical due diligence - e.g. the partner/consultant vetting form; social media research.

B. Do the proposed project partners understand project risks, and are there sufficient risk mitigation or exit strategies in place?

C. Does the project budget ensure that partners are adequately compensated for their contribution to the work?

07
Transparency

A. Have you ensured that consultants and local partners have the information they need to ensure their own safety and wellbeing?

B. Is the service or offer clearly and accurately defined for beneficiaries?

To consider: Does it promise one thing and then provide something else?

08
Accountability

A. What review structures are in place for the project, internal and/or external?

B. How are you ensuring that lessons learned are swiftly captured and addressed?

C. Have you scheduled a project debrief and planned a wrap-up summary?

In addition to the questions above, campaigns and interventions projects must also review the below.

09
Keywords, terms, and content libraries

A. Is the database of keywords and terms coded accurately, using standardised risk rating systems?

> The ethical implications for deploying non-standardised databases is that we end up measuring threats differently in different contexts or across different ideologies. You may also want to consider an independent risk-rater inter-reliability assessment.

B. Is it being developed and updated in a consistent, traceable manner?

> Has time for proactive, research-driven updates been scheduled into workflow?

C. Has the database been quality assured by internal and external subject matter experts (SMEs)?

10
Content curation (including campaign messaging)

A. Who is curating the content?

> Do they have the appropriate skills, expertise and training to effectively and safely meet audience needs? Have you identified suitable SME(s) to help create and/or approve the playlist(s), plus an additional quality assurance step?

B. Have you ensured consent for its use?

> Do you need to think about copyright? What about the safety of individuals who might feature in those videos?

C. Are you comfortable with the origin of the video?

> E.g., has the author created other content that expresses harmful/hateful/violent views, and/or is the author affiliated with an organisation that could be considered problematic?

D. Do you have a comment moderation strategy in place?

> E.g., if you're using YouTube playlists with third party content, is someone monitoring the comments section, and taking action (e.g. removing a video) if the comments become problematic? If content is original, consider whether switching off the comments function is an appropriate safeguarding measure.

11
Content creation

A. Who is creating the content?

> Do they have the appropriate skills, expertise and training to effectively and safely meet audience needs?

B. Has the project found suitable SMEs to design and - separately - review the content?

C. If the project plans to use fictional narratives, such as scripted personal testimonies, is a disclaimer included in the video? If a composite character is created, is this conveyed to the individual?

> Think of how a film might begin with a line such as "this is based on/inspired by real events", etc.

12
Interventions

A. Has the individual voluntarily engaged with the service?

> Is the offer clear (e.g. "I am here to provide mental health support")? If a website is being used, is there a privacy policy visible on the site in a language accessible to users?

B. Does the project envision proactive messaging of individuals on social media/other platforms?

> If so, is there a plan in place to obtain immediate consent to engage and offer an explanation for contact? Who is the messenger - do they have the appropriate skills, expertise and training to effectively and safely meet audience needs?

C. Does the project offer a service (such as counselling)? If so, are the terms and limitations of the service offer clear to the user?

> The risk of over-promising to potentially vulnerable audiences is a key consideration here.