**Responsible Data Curation Checklist for Fairness and Robustness Evaluations**

This checklist translates our HCCV data curation considerations and recommendations into action items for researchers and practitioners. Presented as a series of questions, these items are designed to stimulate discussions among data collection teams. The questions are purposefully worded to avoid binary responses, encouraging open-ended dialogues. The primary focus of the checklist is to underscore the ethical dimensions and ensure that teams address concerns encompassing purpose, consent and privacy, as well as diversity.  
  
It is important to engage with the checklist as a preliminary exercise before beginning data collection. This approach promotes informed decision-making and minimizes risks, leading to more responsible and reliable outcomes.  
  
Contextual diversity is acknowledged to avoid a one-size-fits-all approach. Moreover, customization is encouraged, as not all items apply universally; teams should modify or expand the checklist to align with their context and use case. As with existing AI ethics checklists, it is important to recognize that the checklist is not a guarantee for ethical compliance; rather, it functions as a catalyst for discussion and reflection.  
  
We understand that answering these questions is time-consuming, increasing the burden on data collection teams whose work is already undervalued. Therefore, when navigating through these lists, priority should be put on items related to the specific domain and task of interest. The level of engagement needed for each question will invariably  
differ. Keep in mind that the questions aim to spur ethical thinking during dataset development: “Ethics is often about finding a good or better, but not perfect, answer”.

**Purpose**

The questions in this section focus on eliciting strategies for curating HCCV evaluation datasets specifically for fairness and robustness assessments. They seek alignment with objectives and inquire about factors known to influence these assessments to ensure comprehensive evaluations. Moreover, the questions aim to assist in formulating clear dataset purpose statements, preventing ambiguity and misuse of data, as well as exploring external validation to enhance transparency and accountability.

**Dataset Development Strategy**

* Can you provide details about your strategy for developing a new dataset tailored specifically for conducting fairness and robustness assessments in the context of HCCV? How do you plan to ensure that this dataset is aligned with the objectives of evaluating fairness and robustness?
* Can you elaborate on the factors your dataset will encompass to comprehensively enable fairness and robustness evaluations for HCCV models? How do you intend to capture the primary factors, including data subjects, instruments, and environments, that influence these evaluations?

**Dataset Purpose Statement**

* Can you provide details about your plan to formulate a comprehensive dataset purpose statement? How will this statement effectively communicate the core motivations driving, e.g., data collection, outline the intended dataset composition, specify permissible uses of the data, and identify the specific audience you aim to serve with the dataset?
* Can you elaborate on your strategy for ensuring the accuracy and ethical alignment of your dataset's purpose statement? How do you plan to externally validate the content and ethical considerations of the statement?
* Can you provide insights into the benefits and implications of submitting your dataset's purpose statement as part of a research study proposal in the format of a registered report for your project?

**Consent and Privacy**

The questions in this section explore informed consent, legal compliance, and privacy protection measures within anonymization strategies. The questions emphasize clarity and voluntariness in consent processes to prevent coercion or misuse of data. Moreover, they attempt to elicit strategies for explaining data collection purposes, consent revocation, and accommodating diverse participation circumstances. Furthermore, the questions seek insights into addressing anonymization challenges, aiming to prevent re-identification risks, unauthorized exposure, and legal noncompliance, while preserving data utility and protecting data subjects' rights.

**Informed and Voluntary Consent**

* Can you elaborate on your approach to ensuring that you secure explicit, voluntary, and informed consent from all individuals who either appear in the dataset or can be discerned from it? How do you plan to handle consent for data annotators who may have disclosed personal information for the purposes of quantifying and addressing annotator perspectives and bias?
* Can you provide a comprehensive explanation of your strategy for conveying the purpose of data collection to the subjects? How do you intend to emphasize the utilization of their data, which includes various types of information such as facial, body, biometric images, as well as information about themselves and their environment, all in the context of assessing the fairness and robustness of HCCV systems?
* In what ways will you incorporate consent forms that are composed in plain language to enhance the understanding of AI technologies? How do you plan to make sure these forms effectively convey the intricacies of data usage?
* How do you plan to inform data subjects about their ability to withdraw consent at any given point during, or after, the data collection process? Can you provide details about the mechanisms you will have in place for facilitating this?
* Please provide insight into your strategy for collecting data from individuals below the age of majority or vulnerable individuals. How will you seek both guardian consent and voluntary informed assent in such cases?
* How do you plan to evaluate vulnerability along a continuous spectrum, taking into account contextual factors and recognizing that vulnerability is not solely binary or based solely on group affiliations, but can also be influenced by specific situations or circumstances?
* Can you also provide details about how you will consider the circumstances of participation, which might include the potential need for participatory design, assurances of compensation, provision of educational materials, and safeguards against authoritative structures? How will you address these various aspects in your approach?
* How do you intend to ensure that vulnerable individuals have a comprehensive understanding of the data usage and willingly provide informed assent? Can you outline the specific measures you intend to implement for this purpose?
* Can you elaborate on how you will respect the decision of a vulnerable individual who expresses dissent, regardless of the preferences of their guardian?

**Consent Revocation Mechanisms**

* How do you plan to integrate mechanisms that allow data subjects to easily withdraw their consent? Can you provide specifics on how this process will be designed and executed?
* Can you provide insights into the benefits and implications of implementing dynamic consent mechanisms that utilize personalized communication interfaces? How do you intend to ensure that these mechanisms adapt to the preferences and needs of individual data subjects?
* How do you intend to enable data subjects to actively participate in research activities and manage their consent preferences? Can you provide more details about the tools or processes you plan to put in place to achieve this?
* In what ways will you explore the feasibility of online platforms for consent management that are user-friendly and minimize complexity for data subjects? What steps will you take to ensure easy accessibility?
* Can you provide insights into the options you will provide to data subjects for granting consent? How will you offer choices between blanket consent, case-by-case selection, or opt-in based on specific data usage?
* Can you elaborate on your considerations regarding the formation of a steering board or charitable trust composed of representative subjects from the dataset? How do you envision this entity contributing to decision-making processes?
* How do you plan to empower data subjects to actively participate in decisions concerning the usage of their data? What mechanisms or channels will you establish to facilitate this involvement?
* Can you provide information about the method you will offer data subjects to easily and promptly revoke their consent? How will you ensure that this process is straightforward and accessible?
* How do you intend to address varying levels of technological know-how and internet access among data subjects? Can you detail the measures you will take to accommodate these variations?
* What alternatives do you plan to offer for revoking consent that do not rely solely on online-based processes? How will you ensure that individuals with different needs and preferences can effectively revoke their consent?
* How do you plan to assess the practicality and suitability of the chosen mechanisms for consent revocation, taking into account the expected dataset size and the resources available to you? What criteria will you use to evaluate their effectiveness?

**Country of Residence Information**

* How do you plan to address the fact that anonymization measures might not universally meet legal requirements in specific regions, necessitating additional considerations? Can you provide insights into your strategy for ensuring legal compliance while implementing anonymization?
* Can you elaborate on your approach to collecting information about the country of residence for each individual in your dataset? How do you intend to use this information to ensure legal compliance and address potential privacy concerns?
* How do you plan to familiarize yourself with the data protection laws that are applicable in the countries of residence of your data subjects? Can you provide details about your process for gaining this understanding and how you will apply it to your data curation project?
* How do you intend to prioritize safeguarding data subjects' rights as stipulated by the data protection laws in their respective countries? What steps will you take to ensure that the creation and utilization of the dataset strictly adhere to the relevant data protection regulations? Can you provide specifics about the measures you will put in place to achieve this?
* What mechanisms do you intend to implement to ensure the adaptability of your dataset management strategy to changing legislative requirements? Can you provide details about how you will monitor and accommodate legislative changes in your dataset management approach? Can you provide insights into how you will strike a balance between maintaining compliance and effective dataset management in dynamic legal environments?

**Privacy-Sensitive Image Regions and Metadata**

* How do you plan to implement measures that effectively safeguard against re-identification risks, encompassing singling out, linkability, and inference, within your anonymization approach?
* Can you elaborate on your strategy for redacting all image regions that could inadvertently disclose privacy-related information? How do you intend to comprehensively identify and address these regions?
* Can you elaborate on your strategy for the removal of elements such as body parts, clothing, and accessories for nonconsenting subjects to enhance privacy protection? Can you provide more details about the considerations and methods involved in this process?
* Can you elaborate on your strategy for the removal of text (possibly excluding copyright owner information) from the dataset's images to enhance privacy protection? Can you provide more details about the considerations and methods involved in this process?
* Can you explain your plan for empirically validating the chosen anonymization methods? How will you assess the methods' effectiveness in mitigating re-identification risks while preserving the utility of the data?
* Can you provide details about how human annotators will be engaged in the creation and verification of privacy leaking image region proposals for anonymization purposes? How will you ensure accuracy and consistency in this process?
* Can you provide details about how you intend to align region proposals predicted by algorithms with human judgment, addressing any potential failures or biases? Can you describe your strategy for maintaining a sensitive approach to these factors?
* What steps will you take to address jurisdiction-specific requirements that might necessitate human-generated proposals for biometric identifiers in order to comply with legal and regulatory standards?
* Can you elaborate on the measures you will take to prevent image metadata from inadvertently revealing unauthorized identifying information? How will you ensure that metadata remains privacy-conscious?
* How will you identify specific metadata elements that you intend to retain to ensure a comprehensive understanding during the evaluation process? Can you provide examples of the types of metadata you plan to retain for this purpose?
* How do you plan to replace or remove sensitive information within metadata while retaining its usefulness for fairness and robustness analyses? Can you provide insights into your approach for striking a balance in this regard?

**Diversity**

The questions in this section revolve around obtaining accurate image annotations related to identity, phenotype, environmental factors, and instruments, while upholding inclusivity, sensitivity, and privacy. Additionally, the questions attempt to elicit strategies for documenting identity, ensuring fair compensation, and effective (anonymous) communication.

**Self-Reported Annotations**

* How do you plan to acquire annotations for images directly from the data subjects, leveraging their self-awareness and contextual knowledge to enhance the accuracy and quality of annotations? Can you elaborate on the methods and strategies you intend to use for this purpose?
* Can you elaborate on your strategy for addressing biases and ensuring careful handling when inferring labels about individuals? Can you provide reasoning as to why labels about individuals will be inferred as opposed to being self-identified? How will you actively mitigate potential biases that may arise during the labeling process?
* How do you intend to consider the implications of inferred labels, for example, in relation to data access request rights?

**Versatile and Inclusive Response Options**

* How do you plan to enhance the accuracy and nuance of identity information collection by providing respondents with both closed-ended and open-ended response choices? Can you elaborate on your strategy for using open-ended responses to gather more detailed and comprehensive data?
* How do you intend to ensure inclusivity and prevent any potential implications of exclusion in the response choices you offer?
* Can you elaborate on your preparedness to manage the coding and analysis effort required for processing open-ended responses? What effective strategies do you plan to implement for managing and analyzing the data collected from open-ended questions? How will you handle the potential complexities and variations that can arise from these responses, ensuring that the insights and information derived can be accurately captured and utilized?

**Dynamic Nature of Identity**

* How do you plan to collect self-identified information on a per-image basis, accounting for the fact that identity is intrinsically contextual and temporal? Can you elaborate on your strategy for capturing nonstatic aspects of identity?
* Can you elaborate on your strategy for enabling data subjects to freely choose multiple identity categories without imposing any limitations? How will you ensure that subjects have the flexibility to express their identity in a comprehensive and unrestrictive manner?
* How do you intend to address potential requests for per-image updates to self-identified information provided by subjects over time, respecting their autonomy? What factors have you considered in relation to the potential effects of permitting updates?

**Demographic Information**

* How do you plan to collect precise biological age in years from data subjects to ensure an accurate representation of their age?
* Can you elaborate on your approach to gathering pronoun information from data subjects to enhance gender inclusivity and mitigate the risk of misgendering? How will you ensure that respondents feel comfortable providing this information?
* Can you explain your strategy for gathering consistent ancestry information from data subjects? How will you approach the collection of this information in a sensitive and inclusive manner?
* How do you intend to offer the option for data subjects not to disclose their sensitive attributes if they choose not to? Can you provide more details about how you will handle the sensitivity and privacy of these attributes?

**Sensitive Attributes in Aggregate**

* How do you plan to collect voluntarily disclosed sensitive attributes such as disability and pregnancy status? Can you elaborate on your approach to respecting the willingness of data subjects to provide these details?
* Can you provide insight into your strategy for reporting sensitive attributes, such as disability and pregnancy status, in aggregate data while safeguarding subjects' safety and privacy? How do you intend to ensure that individual identities are protected?
* Can you elaborate on your approach to relying on credible and appropriate sources for the categorization and definitions of sensitive attributes like disability or difficulty? How will you account for the potential variations in these definitions based on cultural, identity, and historical contexts?

**Phenotypic and Neutral Performative Features**

* How do you plan to collect phenotypic attributes, encompassing characteristics such as skin color, eye color, hair type, hair color, height, and weight? Can you provide insights into your strategy for obtaining these attributes in a sensitive and comprehensive manner?
* Can you elaborate on your approach to collecting a diverse range of neutral performative features, including aspects such as facial hair, hairstyle, cosmetics, clothing, and accessories? How do you intend to ensure inclusivity and accuracy in capturing these features?

**Environment and Instrument Details**

* How do you plan to gather data on environment-related factors, which encompass details such as image capture time, season, weather, ambient lighting, scene, geography, camera position, and camera distance? Can you provide insights into your strategy for capturing these factors accurately and comprehensively?
* Can you elaborate on your approach to collecting instrument-related factors concerning the imaging devices used, including aspects such as lens, sensor, stabilization, flash usage, and post-processing software? How do you intend to ensure accuracy in capturing these details?
* How do you plan to obtain environment- and instrument-related information? Can you provide more details about the methods you will use, such as self-reporting, annotator estimation, and sourcing information from Exif metadata? How will you leverage contextual knowledge from image subjects to enhance data quality?
* Can you explain your approach to handling information such as precise geolocation and user-added details in Exif metadata that might contain personally identifying information? How will you ensure compliance with copyright regulations (if applicable) while maintaining privacy and adhering to ethical considerations?

**Annotators as Contributors**

* How do you plan to document the identities of data annotators, including capturing demographic details such as pronouns, age, and ancestry? Can you provide insights into your strategy for gathering and preserving this information while respecting privacy and ensuring transparency?
* Can you elaborate on your approach to highlighting the contributions of annotators beyond data labeling in the dataset documentation after the curation process? How do you intend to accurately represent the multifaceted roles and contributions of annotators?
* How do you plan to report the demographic information of annotators to analyze potential sources of bias in dataset annotations? Can you provide more details about your proposed approach for conducting this analysis while ensuring privacy and ethical considerations?

**Fair Treatment and Compensation**

* How do you plan to ensure that all contributors receive compensation that exceeds the minimum hourly wage of their respective country or jurisdiction of residence? Can you provide insights into your compensation strategy to ensure fair and ethical remuneration?
* Can you elaborate on your approach to exploring alternative payment models, such as compensation based on the average hourly wage? How do you intend to determine a compensation structure that is both fair and reflective of contributors' efforts?
* How will you establish direct communication channels between dataset creators and contributors? Can you provide more details about the methods you intend to implement for effective and transparent communication?
* What communication methods do you plan to explore that maintain the anonymity of contributors? Can you provide insights into your approach to balancing communication and privacy needs, such as using anonymous feedback forms?
* Can you provide information about your strategy for developing clear and accessible plain language guides to facilitate various tasks, such as image submission and data annotation? How do you plan to ensure that these guides effectively assist contributors?
* How do you intend to ensure that contributors from diverse backgrounds can easily understand and follow any instructions provided? Can you elaborate on your approach to promoting inclusivity and accessibility in your communication and guidelines?
* Can you provide details about how you plan to subject your recruitment and compensation procedures to ethics review? What steps will you take to ensure that your procedures align with ethical considerations and best practices?