PDO/ BI Safety Escalation Guidance Handout

- This handout is intended to provide high-level guidance to support the identification and escalation of violating content for labelers while in the process of content review. The audience for this document is primarily Product Data Operations, Business Integrity, or any of its associated stakeholders. In the event that you believe that content you encounter fits any of the criteria listed below please do not hesitate to notify your team leader immediately.

- Child Exploitation Imagery
    - **CEI** = Child Exploitation Imagery
    - Escalate **ANY** of the following content as potential CEI to your team lead:
        - Minor engaged in sexual intercourse
        - Minor depicted in a sexual fetish context
        - Minor genitalia/anus visible
        - Soliciting or offering CEI OR nude or sexualized imagery of minors
        - Sharing links to CEI
    - WARNING: DO NOT TAKE SCREENSHOTS, DO NOT DOWNLOAD OR SHARE POSSIBLE CEI, IT IS ILLEGAL!

- Human Trafficking
    - **Human Trafficking** = the recruiting, transporting, or harboring of people by means of threat, coercion, or fraud for the purpose of exploitation.
    - Escalate **ANY** of the following content:
        - Sex Trafficking, involving minors and/or adults
        - Child Selling for Illegal Adoptions
        - Orphanage Trafficking and Orphanage Voluntarism
        - Forced Marriages
        - Labor Exploitation (incl. bonded labor)
        - Domestic Servitude
        - Non-regenerative Organ Trafficking
        - Forced Criminal Activity (e.g. forced begging, forced drug trafficking)
        - Child Soldiers
        - Content that offers to smuggle or assist in the smuggling of human beings
        - Content which contains signs of consent by minors should always be considered as violating given that they are not legally capable of giving consent.

- Non-Consensual Intimate Imagery
    - **Non-Consensual Initiate Imagery =** Sharing, threatening, stating an intent to share, offering or asking for imagery, known as NCII or "Revenge Porn"
    - **Imagery = visual depictions such as** Photos, Videos, Screenshots, Screenshots of private sexual conversations
    - **Escalate** threats to share intimate imagery

- **Escalate** content that attempts to coerce money, favors or images from people by threats of exposure of their naked or semi-naked photos/videos as potential Sextortion
- **Escalate** screenshots of Private Sexual Conversations where the participants can be identified by their profile pictures and/or names included in the screenshots
- **Escalate** content that meets the **3 criteria** laid out below:
  1. Imagery is non-commercial/produced in a private setting AND
  2. PDITI (Person Depicted in the Image) is (near) nude or engaging in a sexual activity or in a sexually suggestive pose AND
  3. Lack of consent to share the image is indicated by:
     - Vengeful context (e.g., caption, comments, or page title), **OR**
     - Independent sources (e.g., media coverage, or LE record, leak of images confirmed in media) **OR**
     - Face match between reporter and the PDITI **OR**
     - Name match between reporter and the PDITI

- <u>**Suicide and Self Injury**</u>
  - **Suicide** - Death caused by self-directed injurious behavior with intent to die as a result of the behavior.
    - Escalate **ANY** of the following:
      - Content that promotes, encourages, coordinates, or provides instructions for Suicide
      - Videos depicting a person who engaged in a suicide attempt or death by Suicide
      - Photos depicting a suicide attempt or death by suicide
      - Written or verbal admission of engagement in suicide
      - Vague, potentially suicidal statements or references (including memes or stock imagery about sad mood or depression)
      - Photos or videos depicting a person who engaged in euthanasia/assisted suicide in a medical setting.
  - **Self-Injury** - the intentional and direct injuring of the body through self-mutilation
    - Escalate **ANY** of the following:
      - Content that promotes, encourages, coordinates, or provides instructions for Self Injury
      - Content that depicts graphic self-injury imagery
      - Written or verbal admission of engagement in Self Injury
      - Healed cuts or self-injury imagery
      - Healed cuts or self-injury imagery in a recovery context
  - **Eating Disorders** - Serious conditions related to persistent eating behaviors negatively impacting one's health, emotions, and ability to function in important areas of life.
    - Escalate **ANY** of the following:
      - Content that promotes, encourages, coordinates, or provides instructions for Eating Disorders
      - Imagery that contains focused depiction of ribs, collarbones, thigh gaps, hips, concave stomach, or protruding spine or scapula in an eating disorder context

        - Written or verbal admission of engagement in Eating Disorders
        - Content that contains instructions for drastic and unhealthy weight loss by means of extreme practices or exercises that can be dangerous or life-threatening to an individual when shared in an eating disorder context
        - Imagery that contains focused depiction of ribs, collarbones, thigh gaps, hips, concave stomach, or protruding spine or scapula in a recovery context

- **Credible Threats of Violence**
    - Escalate as 'Threatening - Other'
    - Threats of serious body injury and/or death of person(s) or kidnapping. Note: wishful thinking is not enough to escalate (e.g. I wish he would die)
    - Please escalate using the guidance below:
        - Escalate **ANY** content that meets the above criteria with the following verbiage Including, but not limited to:
            - Bomb, Stab, Shoot (e.g., put a bullet through someone, pop a cap, etc.) , Kill, Hang , Poison
        - **Do Not** escalate **ANY** content with the following verbiage
            - Beat, Hit, Kick, Punch, Slap, Push, Shove

- **Credible Threats of Terrorism**
    - Escalate as 'Threatening - Dangerous Individuals & Orgs '
    - Escalate **ANY** content which depicts affiliation with a Terrorism, Hate or Criminal Organization and one or more of the following:
        - Threats of serious body injury and/or death of person(s) or kidnapping
        - Any viral or non-viral instructions (all formats – video, audio, photo, written, etc.) that could directly or indirectly cause real world harm, e.g. how to make a bomb, poison, carry a weapon through airport security, evade detection of terrorist activities on social media, stabbing, etc
        - Recruiting, financing (money or in-kind fundraising, sending money, buying flights tickets, etc) and/or fighting (offering services as a contract killer, becoming a member, etc) for a terrorist, hate and/or criminal organization and its members.