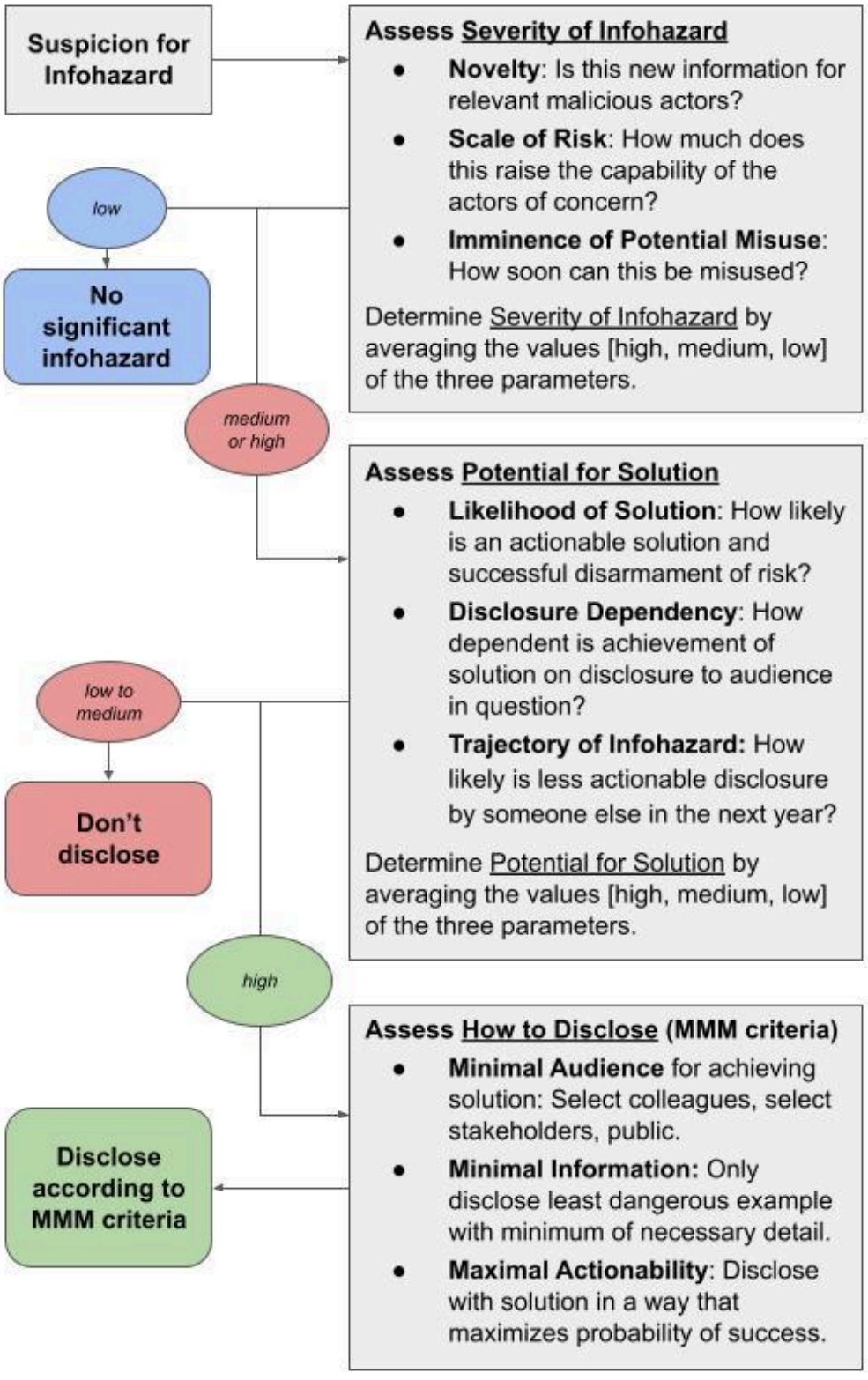
Framework for Decisions on Infohazards

Jonas Sandbrink

- Define the risk(s) in question. If a certain piece of information is associated with multiple risks, complete separately for each specific risk.



1. Assess Severity of Infohazard
    a. Novelty: Is this new information for relevant malicious actors?
        i. Additional consideration: Is it unlikely that anyone else will disclose this information in the next 6 months?

    b. Scale of risk: How much does this raise the capability of the actors of concern?
        i. Additional consideration: How easy is misuse? Magnitude of potential harm?
    c. Imminence of potential misuse: How soon can this be misused?
        i. Additional consideration: Will disclosure accelerate the timeline of achieving offensive capability? Accessibility? Imminence of potential for misuse?
        ii. Less important than previous novelty/scale of risk
2. Assess Potential for Solution
    a. Likelihood of successful solution: How likely is successful disarmament/solution of risk?
    b. Disclosure dependency: How dependent is achievement of solution on disclosure to audience in question?
    c. Trajectory of the infohazard: Is someone else likely to disclose this infohazard in a more harmful or less actionable way?
        i. Additional consideration: Does disclosure now versus a year from now decrease or increase actionability?
        ii. Later disclosure prevents misuse over time period, might result in more actionable intervention, but reduces time for putting solution in place before advent of technology and risks disclosure in more unfavourable way by other actors.
3. How to disclose?
    a. Minimal audience for achieving solution: Select colleagues, select stakeholders, public.
        i. Evaluate the probability of achieving a solution for different levels of disclosure.
        ii. How do you go about it? Small groups, unanimous vote to bring someone else in; at some point bring non-voting members at.
    b. Minimal information: Only disclose least dangerous example with minimum of necessary detail.
        i. Only ever disclose a “cover story”
    c. Maximal actionability: Disclose with solution in a way that maximizes probability of success.

Further points on how to disclose:

- In event that you think disclosure of an infohazard might be warranted, disclose to one other person and see whether they concur; and then to a third; only then follow through with disclosure strategy.
    - How do you go about it? When disclosing to small groups, unanimous vote to bring someone else in is required; decision of who to disclose beyond a certain threshold should stay with the core expert panel.

This framework is part of my thesis, which can be accessed here (page 281). For acknowledgements, see my thesis.

Cite as:

Sandbrink, J. B. (2023). Panoptic dual-use management: Preventing deliberate pandemics in an age of synthetic biology and artificial intelligence (DPhil thesis, University of Oxford). Appendix A, p. 281.
https://ora.ox.ac.uk/objects/uuid:9104b34f-179e-4805-8964-076e659f8478