Seeking Justice & Remediating Harms

Betsy Popken, Executive Director, UC Berkeley Human Rights Center

Using LLMs at Work

- Who here uses ChatGPT?
- How many of you use it in your work?
- What work do you do? / How do you use it in your work

UN Guiding Principles on Business & Human Rights

“In order to meet their responsibility to respect human rights, business enterprises should have in place policies and processes appropriate to their size and circumstances, including…processes to enable the remediation of any adverse human rights impacts they cause or to which they contribute.”

Human Rights Assessment of LLMs

- Literature review
- Global stakeholder interviews
- Assessment of human rights risks
- Recommendations

Human Rights Risks

- Journalism - access to information & downstream risks
- Law - due process and right to a fair trial
- Education - freedom of thought

Recommendations to Remediate

- Provide filters (display text with warning or not at all)
- Consult experts and users in different parts of the world
- Assess potential risks of use in high-risk professional contexts
- Provide opportunities for users to give feedback
- Educate users to not take the output at face value
- Focus on harm to humans when looking at risks and AI principles (Who are the most at-risk communities?)
- Conduct risk assessment along different parts of the AI lifecycle (datasets, model outputs)

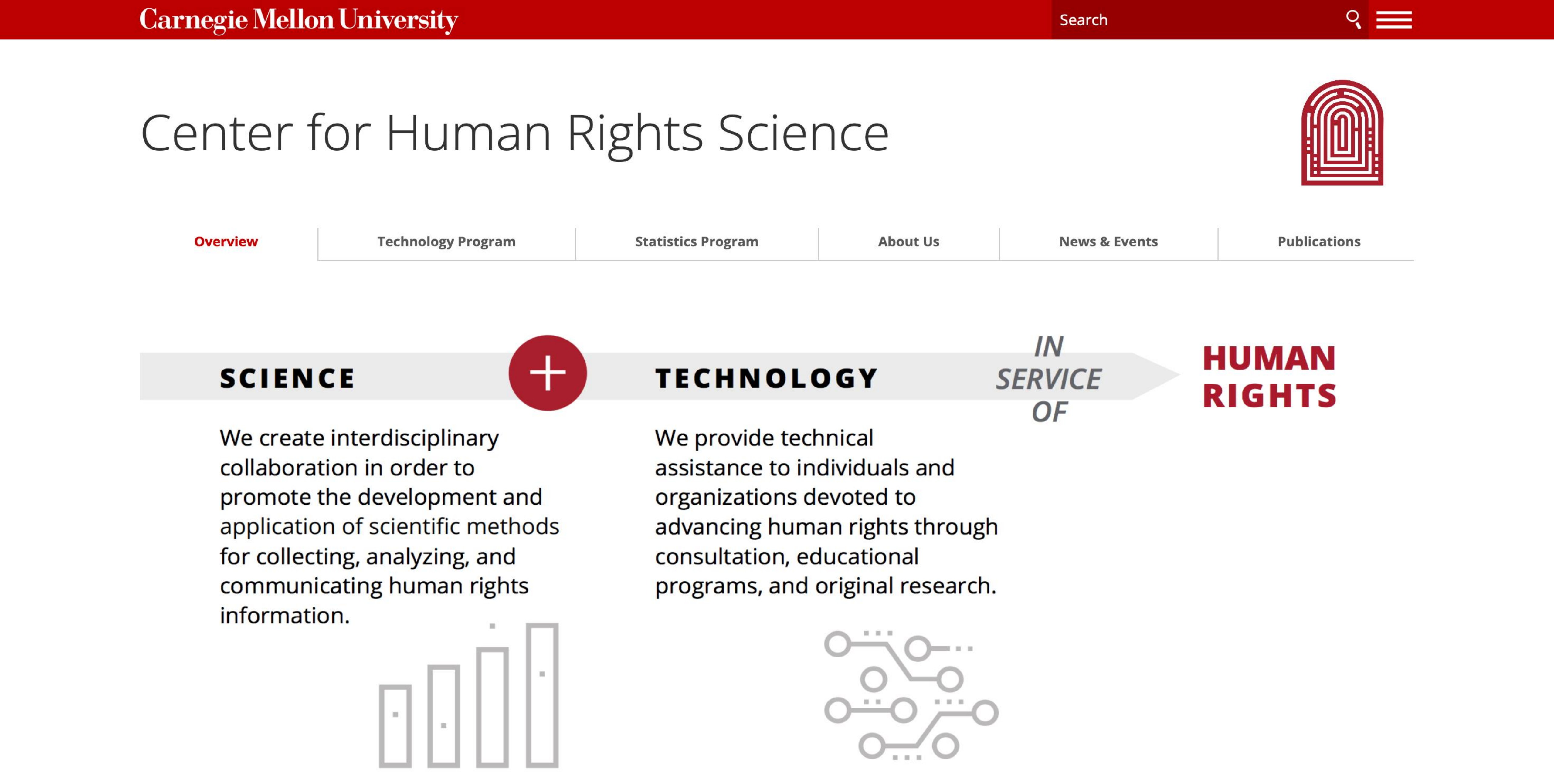
Thank you!

Carnegie Mellon University
Center for Human Rights Science

Technology, Human Rights, and Harm Reduction

Jay D. Aronson

November 19, 2024



Center for Human Rights Science

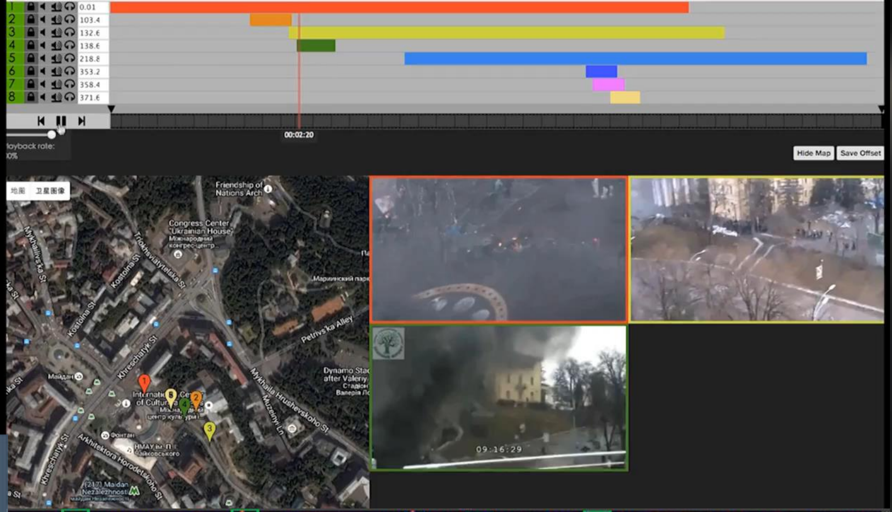
SCIENCE + TECHNOLOGY IN SERVICE OF HUMAN RIGHTS

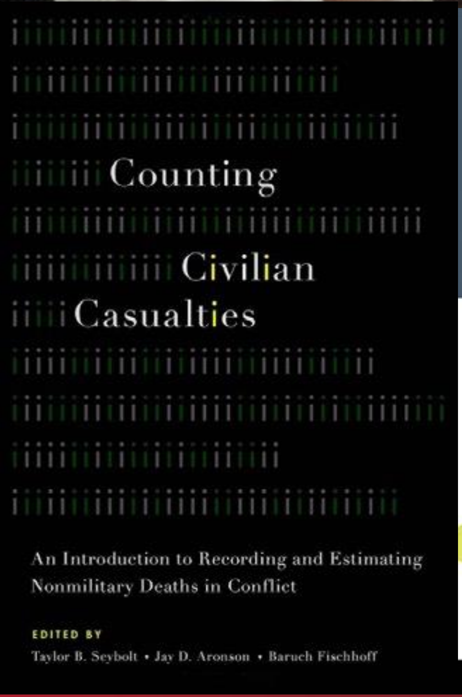
SCIENCE

We create interdisciplinary collaboration in order to promote the development and application of scientific methods for collecting, analyzing, and communicating human rights information.

TECHNOLOGY

We provide technical assistance to individuals and organizations devoted to advancing human rights through consultation, educational programs, and original research.

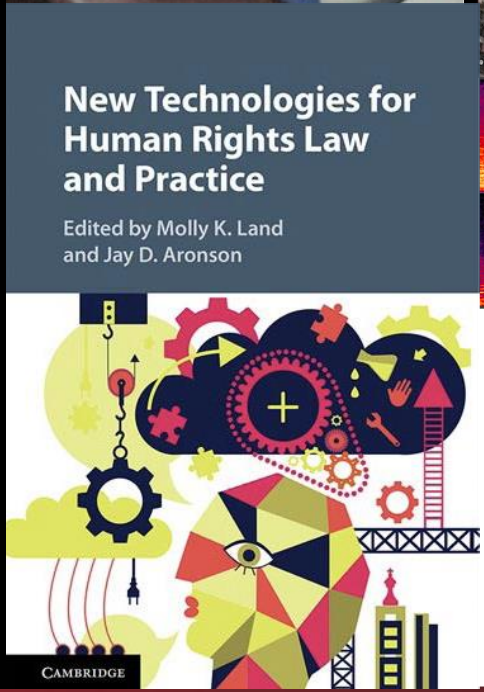




Counting Civilian Casualties

An Introduction to Recording and Estimating Nonmilitary Deaths in Conflict

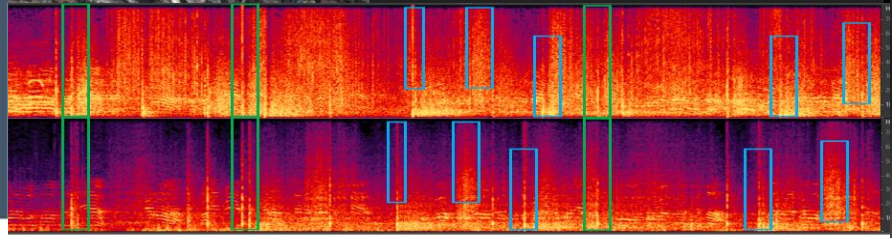
EDITED BY
Taylor B. Seybolt • Jay D. Aronson • Baruch Fischhoff

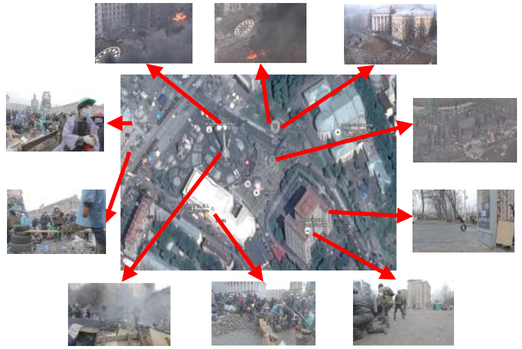


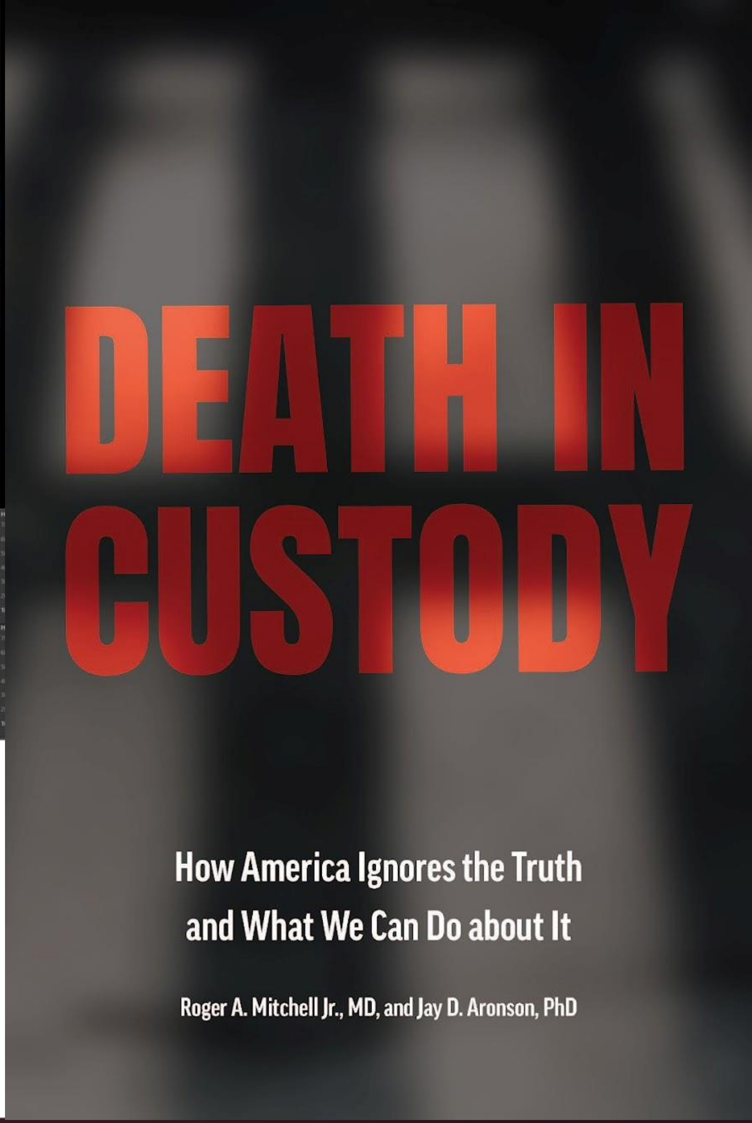
New Technologies for Human Rights Law and Practice

Edited by Molly K. Land and Jay D. Aronson

CAMBRIDGE







DEATH IN CUSTODY

How America Ignores the Truth and What We Can Do about It

Roger A. Mitchell Jr., MD, and Jay D. Aronson, PhD

Center for Human Rights Science

Core Values: How to Limit Harm

- Collaboration
- Solidarity
- Humility
- Tech skepticism
- Trust
- Deliberate/Cautious Action

Cliché but True: Double-Edged Sword

- Technologies that can promote and protect human rights can also be used to violate human rights
- Very difficult to have the positives without negatives
- Features that make technologies useful also make them dangerous
- Technology isn’t neutral! Design reflects particular norms and values

Human Rights Approach

- Human rights law as a source of norms and practices
    - Focus on most vulnerable/least powerful
    - Enhance ability to make seek accountability and make rights claims
    - People impacted by tech must be meaningfully involved – not just consulted – in development and design
    - Recognition of power and privilege built into technological systems →tech must be engineered to have positive impact; it doesn't just happen magically
    - Recognition that tech impacts all aspects of our lives and rights

Thank you!

WWW.CMU.EDU/CHRS

Carnegie Mellon University
Center for Human Rights Science

240 BAKER HALL
5000 FORBES AVENUE
PITTSBURGH, PA 15213
CHRS@CMU.EDU

Oversight Board

Seeking Justice and Remediating Human Rights Harms

NASEM Symposium: Engineering and Human Rights, November 19, 2024

INDEPENDENT EXPERTS FROM AROUND THE WORLD

Committed to free expression and human rights

STRATEGIC PRIORITIES

Elections &
Civic Space

Crisis & Conflict
Situations

Gender

Hate Speech Against
Marginalized Groups

Government Use
of Meta's Platforms

Treating Users
Fairly

Automation Enforcement
of Policies and Curation
of Content

CASES AND RECOMMENDATIONS

- Appeals from users to remove or restore content
- Referrals from Meta
- Recommendations on
  - policy
  - transparency
  - enforcement

CASE EXAMPLE

Raising awareness of Breast Cancer

Our recommendation: “improve the automated detection of images with text-overlay to ensure that posts raising awareness of breast cancer are not wrongly flagged for review.”

- Ensuring content goes to human review vs. automatic renewal
- Improvements to **Meta’s** text-overlay detection led to 2,500 pieces of content being sent to human review over just a 30 day period in February and March 2023
- Led to testing & deployment of a new health content classifier for identifying image-based content about breast cancer.
- **In one month** an additional 1,000 pieces of content being sent for human review that would have previously been removed without review.
- Systemic changes and new tools – and shows the different ways in which Meta is responding to this important recommendation.

CASE CRITERIA

Difficulty

Disagreement
Is there strong disagreement on whether a piece of content should remain on Facebook, IG or Threads?

Clarity
Are Meta's Community Standards clearly articulated?

Consistency
Are Meta's Community Standards applied consistently to this type of content?

International Human Rights Concerns
Are Meta's Community Standards regarding this type of content consistent with international human rights principles?

Significance

Severity
Does the decision to leave-up or take-down the content have severe consequences that could impact users?

Diversity
Does this case add to linguistic or geographic diversity?

Scale
Did a large number of people see, engage with, or react to the content?
Does the content illustrate a larger trend or issue on the platform?

Value to Public Discourse
Has the content been a topic of regional, national, or international discussion?
Does the content have outsized value to public discourse?

PUBLIC COMMENTS

As part of our decisions process, individuals and organizations can **submit public comments**

This input is crucial to achieving our goal of improving how Meta treats people and communities around the world. On numerous occasions, public comments have **shaped our decisions and recommendations** to Meta.

We welcome comments from individuals and organizations around the world.

Oversight Board

THANK YOU

engagement@osbadmin.com

www.oversightboard.com

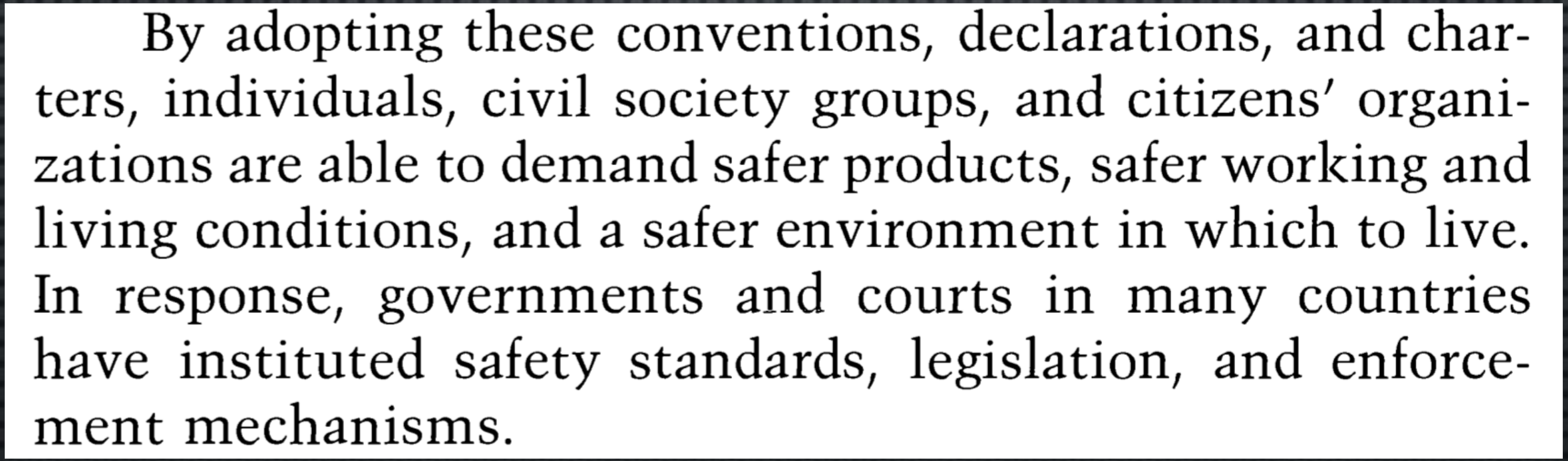
August 20, 2024

Science Supporting Safety as a Human Right

José L. Torero, University College London, United Kingdom

People's Right to Safety

- On 10 December 1948, the General Assembly of the United Nations (UN) adopted and proclaimed the Universal Declaration of Human Rights (UDHR). Article 3 of this Declaration states, "Everyone has the right to life, liberty and security of person."



By adopting these conventions, declarations, and charters, individuals, civil society groups, and citizens' organizations are able to demand safer products, safer working and living conditions, and a safer environment in which to live. In response, governments and courts in many countries have instituted safety standards, legislation, and enforcement mechanisms.

Mohan, D., 2002. Safety as a human right. Health & Hum. Rts., 6, p.161.

- What is the role of science in supporting Safety as a Human Right?

THE SCIENCE OF SAFETY?

- Safety many times can be "Common Sense" but most of the times is a complex technical problem that requires high levels of competency from all those involved
- What is your responsibility when you hold that knowledge and competency, but you are being A witness to the mis-use of science and technical knowledge to the detriment of people's safety?

MAINTAINING A TECHNICAL ADVISORY ROLE WHEN FACING INJUSTICE

- Chile: Governments poor management practices vs. Those who apply them
- Paraguay: Obsolete regulatory practices vs. Building Manager
- Afghanistan: Ignorant practices of Aid Organizations vs. The unqualified user
- South Africa: Irresponsible Corporate Practices vs. The employee
- Peru: Incompetent application of International regulation vs. The operator
- Mexico: Inappropriate use of technical knowledge vs. The bereaved
- United Kingdom: Government Regulation promoting social inequity vs. The Vulnerable