HARPOCRATES

An information hazard mitigating online publishing system unreliant upon censorship or silos

Archimedes Network

The Problem:

Peer-review is sometimes proposed as a soft-censorship backstop against publishing information hazards. However, this fails in several ways:

Censorship fails at preventing the spread of most information hazards: Secrets in empirical disciplines (including all sciences and all security studies) are subject to independent discovery making them functionally impossible to keep.

Censorship induces self-defeating incentives: Censorship, whether from the decisions of an editorial board, peer-reviewers, or self-censorship, only suppresses information hazards from conscientious authors. Less responsible authors do not self-censor, and easily reach prominence in their fields over the conscientious ones by bypassing censors to use uncontrolled publication channels. Also, specific details of dangerous information must be shared to protect against the harm that it represents. Therefore, even responsible authors have incentive to bypass controls over information hazards.

Silos damage research quality in security: Peer-reviewed journals are formulated for readers in narrowly defined academic fields. Thus, leaving information hazard management in their hands has the unintended consequence of siloing security studies and thus sabotaging its utility. This is because, unlike the academic fields they are attached to, security studies are intrinsically multidisciplinary dealing with all factors related to offense, defense, and consequences surrounding a particular class of threat.

Speed matters in security publishing: Security studies inform policy; crises drive security policy, and crises change on a time frame of hours or days, not months or years like the publication time of most journals. This means a preprint-like near instantaneous publishing model is required, particularly for information hazard laden results..

Information Hazard refers to the risk of harm from the dissemination of a true piece of information. For example, disseminating information on bomb making, could lead to otherwise avoidable bombings. The conundrum lies in the fact that information on bomb making is also needed for bomb detection.

Security Studies are academic fields that study adversarial security of various aspects of human civilization (e. g. biosecurity). Information hazards are intrinsic to all security studies.

Digital rights management (DRM) is a series of technologies that try to control redistribution of electronically published material. They are widely considered failures because they reduce audience engagement, and yet don't prevent or even long delay the pirating of protected content.

Preprint services distribute academic work online without a peer-review process. Beyond screening for spam, they make no attempts at validation of the *content* of the studies they publish. They have been widely adopted because they are much faster than traditional peer-reviewed journals.

The Solution:

Named "Harpocrates" after the Greek god of secrets and hope, the publishing system would initially be tested as a preprint service. At the same step as the service screens submissions for spam, it would also screen submissions for information hazards using a mixture of vetted volunteer readers and AI. All non-spam submissions would still be published promptly. However, information hazard containing work would also be subject to one or several DRM interventions based upon how severe the perceived information hazard risk is.

For this use of DRM, the reduction of audience engagement which is typically perceived as a downside, is *the primary feature*. Unlike censorship, which creates security-defeating incentives for the field, this combination of prompt but throttled publication creates a virtuous cycle of incentives that reduces exposure of the public to information hazards, reduces their severity, and enhances the quality of security research:

- Authors have incentive to use Harpocrates to demonstrate their due-diligence in protecting the public against information hazards to peers, funding entities, and tenure committees.
- Authors also have incentive to work with Harpocrates to voluntarily alter their text to mitigate information hazard exposure in exchange for a less intrusive DRM intervention.
- No longer tied to narrow academic fields, security studies lose their silos, and publish faster.

Over time, the Harpocrates method could be generalized as a standard for all sorts of online publishing. Preprint services for security studies are merely the lowest hanging fruit.