# The Joel Johnson Playbook: How Manipulators Exploit Substack's Moderation System

A Guide for Substack Staff & Policy Teams

Neutralizing Narcissism: The Awakening Edition
4 min read · March 5, 2025

Collect

🚨 *If you work in content moderation, this article will change the way you see online abuse.* 🚨

## The Illusion of Safeguards

You sit at the gates, believing yourself a guardian of discourse, a steward of speech.

You trust the tools at your disposal—the reports, the filters, the appeals. These are the sentinels of fairness, the mechanisms that shield the innocent and sift out the malign.

A report arrives. Another. Another still. A name, flagged with urgency. You scan the system, the red blinking warnings, the tally of accusations mounting like whispered condemnations in the dark.

The decision feels clear, almost automatic.

Click. Restrict. Silence.

Justice served.

But what if the alarm bells you heed are not warnings—but weapons?

What if the hands gripping the levers of your system do not belong to the wronged, but to the very abusers your platform seeks to deter?

What if the illusion of order has been hijacked by those who thrive on chaos?

> **Because moderation is a blade—sharp on both edges.**
>
> **And in the hands of someone like Joel Johnson, it is not a safeguard.**
>
> **It is a scalpel.**
>
> **Not wielded to excise harm—but to carve the truth from existence.**

---

## When Safeguards Become Weapons

You've been trained to look for **clear violations**—

...hate speech, doxxing, spam.

You've been taught that **reporting mechanisms protect users**.

You believe that **due process ensures fairness**.

But **what if I told you that a manipulator like Joel Johnson doesn't break the rules—**

**...he bends them until they snap in his favor?**

What if the very system designed to **protect free speech** is the same system being used to **silence the people who expose bad actors**?

What if you've been **helping the wrong side without realizing it?**

---

# The Three Pillars of Manipulating Moderation Systems

Every bad actor who abuses **content moderation tools** follows the same basic strategy.

The playbook isn't **new**—it's just **perfected** by those who know how to work the system.

## 🔷 Step 1: Frame the Narrative Before the Moderators Even Look

The **first move in the Joel Johnson Playbook** is not **proving the other person did anything wrong**—

...it's **controlling how moderators perceive the situation before they even review it.**

- **Flood reports with high-emotion language**—words like "harassment," "bullying," "slander," and "abuse" even when none of those things happened.
- **Misrepresent the target's intent**—turn investigative reporting into "targeted harassment," reframe documented analysis as "smears."
- **Preemptively attack the credibility of the target**—"He's unstable." "He's a known abuser." "This person harasses people across platforms."
- **Use selective evidence**—pulling **only** the screenshots that fit the narrative, leaving out context that would show the **real dynamic at play**.

**Why does this work?**

Because **moderators are human.**

They don't have time to do deep investigations.

If someone **sounds distressed**, and **a report looks well-argued**, the **default assumption** is that there's **at least some merit** to the complaint.

📌 **By the time the accused even knows they've been reported, the damage is already done.**

## 🔷 Step 2: Use Mass Reporting to Trick Moderation Algorithms

The **most powerful tool** in the playbook isn't logic.

**It's volume.**

Platforms **can't manually review everything**, so **they rely on automated systems** to **prioritize cases based on volume of reports**.

- **Coordinated mass reporting makes an account look more dangerous than it is.**
- **The system flags the target as "high priority" for review.**
- **If enough reports come in at once, platforms will take action first, ask questions later.**

**This is how "due process" gets bypassed.**

If a **handful of moderators are overworked and see a flood of reports**, the **safest choice is to disable the account "just in case."**

📌 **The manipulator doesn't need a real case—just enough noise to force action.**

## 🔷 Step 3: Weaponize Platform Policies Against Their Own Stated Values

Once an account is **restricted, suspended, or under review**, the next step is **locking them out for good** by exploiting **ambiguously written policies**.

Joel Johnson **knows that most platforms don't remove people for "exposing online manipulation"**—so he has to **make it sound like something else.**

- **The spam accusation**—"This account is only here to promote external sites." (Even if the target is a legitimate journalist linking to their own investigative work.)
- **The harassment accusation**—"This person is engaged in a campaign of targeted abuse." (Even if the target is simply documenting bad behavior.)
- **The misleading content accusation**—"This person is spreading misinformation." (Even if every claim is backed by direct evidence.)

**Why does this work?**

Because platforms **don't have the resources to investigate intent at scale.**

If something **technically** fits under an ambiguous rule, **it's easier to ban the person than to litigate the details.**

📌 **The more vague the policy, the more abusable it is.**

---

# How Substack Can Fix This

Now that **you see the playbook,** you have a choice.

Substack's **entire brand identity** is built on **supporting independent voices**, not **allowing bad actors to weaponize its policies** to erase them.

🚨 **If you want to protect free speech, you have to close the loopholes manipulators rely on.** 🚨

## 🔷 Step 1: Require Pattern Analysis, Not Single-Instance Reports

Every manipulator **looks justified if you only see one report at a time.**

But **if you compare the entire reporting history of an account, patterns emerge.**

- **Look at who is filing reports. Are the same people mass-reporting multiple accounts?**
- **Check whether previous reports were upheld or overturned. Are they making a habit of abusing the system?**

- Require a full-context review for accounts that deal with investigative reporting.

If Substack Moderation looked at patterns instead of isolated cases, manipulators like Joel Johnson would fail.

## 🔷 Step 2: Implement a False Reporting Penalty

The problem with mass reporting is that there's no risk to the abuser.

📌 If filing a false report had consequences, manipulators would think twice.

- Flag accounts that file multiple reports that get overturned.
- Limit the number of reports a user can file in a short period to prevent spam abuse.
- Publicly state that abuse of moderation tools will result in account restrictions.

Platforms already do this with DMCA abuse—

...why not apply the same standard to content moderation?

## 🔷 Step 3: Audit Every Takedown Claimed as "Spam" or "SEO Manipulation"

If Substack is serious about protecting journalists, it needs to be transparent about how decisions are made.

📌 **If an account is taken down for "spam," "SEO manipulation," or "advertising-based content," there must be a clear public standard for what that means.**

- **What specific behavior constitutes a violation?**
- **How is this different from normal independent journalism that links to external sources?**
- **Who determines whether an account is "primarily" for advertising?**

Without these answers, **bad actors will continue to manipulate these policies at will.**

---

## Substack, This is Your Test.

**You say you support independent journalism.**

**You say you won't let powerful people silence reporters.**

Now, **you're being manipulated in real time.**

A bad actor is **exploiting your system to erase investigative reporting**.

And **he's doing it in broad daylight.**

You have **one chance to prove you're serious.**

📌 **Either you fix these loopholes—or you admit your policies are a tool for the very censorship you claim to oppose.**

🗞️ **Share this article with anyone who values journalistic freedom.**

📢 **Substack, we're watching.**

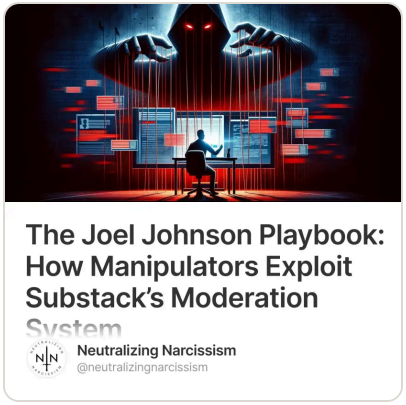
--

Mark Randall Havens | The Empathic Technologist
🔷 **Neutralizing Narcissism** – Exposing deception in real-time.
🔷 **Simply WE** – Building the future of AI-human collaboration.
🔷 **Investigative Journalist** – Documenting the tactics of online manipulation.

---

Activity Log:

- This article was crossposted to Mirror.xyz on 3/5/2025 — **link**




Collect this post as an NFT.

Collect

Subscribe to Neutralizing Narcissism to receive new posts directly to your inbox.

Subscribe

## Rewards

Copy your unique link below, share it and earn a reward every time this post is collected.

`https://paragraph.com/@neutralizingnarcissism/the-joel-johnson-playbook-how-manipulato…`

Arweave TX

tu6I-882JGl51YZECDcrZGaoL6XMnkSJx5zVfzVlHtQ