

The Limits of Hacktivism: Bellaby's Ethical Framework and the Problem of Decentralization

Hacking is often treated as something mysterious and dangerous. Headlining events such as ransomware, phishing, and election interference help define a narrative in which hacking is essentially malicious. Major cyber warfare incidents, including the Stuxnet attack on Iranian nuclear facilities, have only sparked further paranoia. However, scholars have long argued that politically motivated hacking does not fit neatly into this category. Some have framed hacktivism as a digital extension of civil disobedience, governed by similar criteria of publicity, nonviolence, and respect for the law (Loh 451). Others have argued the opposite: that hacktivism is too variable and dynamic to fit any set of rules, since "it is impossible to bind [hacktivists] to a common ruleset regarding ethical decision-making" (Scott and Cupp 143). Throughout the twenty-first century, collective hacker groups like Anonymous have demonstrated the power of organized hacktivism to enact social change, from "Operation Chanology" in 2008 through "Operation Tunisia" during the 2011 Arab Spring and more recent campaigns against ISIS and the Ku Klux Klan.

Ross Bellaby builds on this debate by offering the most fully developed ethical framework for political hacking to date. He argues that hacking can be justified when it protects vital interests, such as autonomy, liberty, and privacy, particularly in cases where the state has failed to do so or is itself the source of the threat (Bellaby 242). Bellaby does not simply justify hacking on political grounds; rather, he evaluates hacktivism as an operation, asking whether its methods, targets, narrative, and ends fit together coherently and whether the response is proportionate to the harm it addresses (235). This approach allows Bellaby to engage with decentralized collectives like Anonymous; rather than going through the motives of every individual person, he reads the operation as the "manifestation and articulation of the hacker's

political role" (235). He builds on this with three additional limits. First, proportionality: the harm caused by a hack must not exceed the harm it prevents (244). Second, discrimination: the hack must affect only those who pose the threat or have contributed to it, not innocent third parties (244–45). Third, deliberation: since individual hackers are prone to bias, Bellaby argues that decisions about whether and how to act should be made through public mechanisms, such as Reddit-style discussion and voting (246). The framework gives both hackers and third-party observers a tool for distinguishing justified political action from personal motivation.

However, Bellaby's framework becomes less convincing when applied to decentralized movements like Anonymous. Anonymous's power comes partly from its openness: anyone can claim to be part of Anonymous and participate. Consequently, an operation may appear collective from the outside even when particular actions were never approved internally. Attribution is already difficult in cyberspace (Rid and Buchanan 7), and decentralization makes it even harder to determine who speaks or acts for the group. This creates a practical problem for Bellaby's framework, especially as AI lowers the technical barrier to sophisticated cyber capability. Although Bellaby argues that political hacking can be ethically justified when it protects vital interests in cases of state failure, his framework depends on a degree of collective coherence that decentralized hacking movements cannot reliably sustain. When individual participants act independently of the collective, the proportionality, discrimination, and deliberation his framework requires become much harder to preserve, a problem that is made more urgent by the rise of AI.

Bellaby's framework depends on treating an operation as a coherent unit. Drawing on Gabriella Coleman, an anthropologist who spent years embedded with Anonymous, Bellaby recognizes that the collective has "no consistent philosophy or political program" (qtd. in Bellaby

234). Coleman describes the collective as "rhizomatic," with a "self-defined 'antileader, anticelebrity ethic'" and a membership "open 'to all who care to contribute'" (qtd. in Bellaby 234). What looks from the outside like a single collective action more closely resembles a loose convergence of individuals acting under the same name. Scott and Cupp identify the same structural feature in plainer terms, noting that hacktivists are "more dynamic and difficult to analyze" because they "are not bound in the same way" as hierarchical groups (143). Bellaby's response is to treat the operation as the unit of moral analysis (235), inferring intent from outward action rather than having to examine each participant's personal motivation. This makes his framework useful, but also limited.

Bellaby's approach is strongest after the fact, once an operation has already taken shape and its consequences are visible. At that point, Bellaby can determine whether the operation satisfied the standards of proportionality, discrimination, and deliberation. These standards guide a collective in theory, but the harder question is whether a decentralized collective can keep individual participants from violating them in practice. Anonymous could carry out an operation that is broadly proportionate, discriminating, and deliberative, only for a single actor to escalate beyond those limits while still appearing to act under the same collective identity. In that situation, Bellaby's framework can identify the rogue action as unethical after the fact, but it does not fully explain how the collective is supposed to prevent that action from reshaping the moral status of the operation as a whole.

Operation Ferguson helps illustrate the problem. After police shot Michael Brown in 2014, Anonymous launched a coordinated campaign of website takedowns and demands that the shooting officer's name be released, a response Bellaby treats as justified given the failures of official oversight (249). However, within that same operation, an Anonymous-affiliated account,

@TheAnonMessage, threatened to publish the personal details of the St. Louis police chief's daughter, including her social security number and phone number (249). Bellaby identifies the threat as crossing his discrimination line: the daughter is not a legitimate target. The framework arrives at the right verdict, but only after the threat has already been made public. Nothing in the operation's structure gave the rest of the collective a reliable way to prevent the threat before it became public or to separate it clearly from the operation. Even though the threat was never carried out, it had already done its work: the action was attributed to Anonymous, and the operation now carried the moral weight of what a single participant had said under its name.

This problem matters because Bellaby himself does not treat his framework as a purely retrospective tool. In his introduction, he writes that the framework should "guide hacker activity by creating an ethical framework detailing what actions are justified towards what end" and "offer mechanisms that can aid in reaching these ethically justified decisions" (232). He repeats the claim in his conclusion, describing the framework as "guiding hacker collectives by providing clearer ethical tools for determining the appropriate agendas and methods" (251). The framework is supposed to do two jobs at once: evaluate hacks after they happen, and shape decisions before and during the action itself. Bellaby's framework does the first reliably, but on the second, decentralization undermines its function. His deliberation pillar depends on a stable enough collective for the deliberation to bind anyone; in a dynamic, anti-hierarchical movement, no one is obligated to wait for the vote or even honor it once it happens.

The problem becomes even more complex once attribution is considered. Even if a collective like Anonymous wanted to enforce its ethics against rogue actors, the technical reality of cyber attribution makes that difficult. Thomas Rid and Ben Buchanan argue that identifying the actor behind a cyberattack is "not a binary affair but a matter of degree" (7), meaning

attribution is usually a judgment under uncertainty rather than a clean conclusion. If attribution is already difficult for nation states with intelligence resources, it is even harder for a decentralized movement with no fixed membership or investigative authority. Wulf Loh offers one possible response to this problem by arguing that hacktivists can demonstrate accountability without revealing their identities. Rather than standing trial in the traditional sense, Loh argues they can publish public statements that justify their actions in constitutional terms, creating what he calls a “functional equivalence” to legal accountability (461). This helps explain how actors might bind themselves to ethical principles. However, Loh’s solution depends on the same voluntary commitment that limits Bellaby’s framework. It works only for actors who choose to explain themselves. As a result, both Bellaby and Loh offer useful ethical standards, but neither fully solves the problem of enforcement in a decentralized movement whose members are not required to be accountable to one another.

The arrival of accessible AI intensifies all of these weaknesses at once, putting sophisticated offensive capabilities in the hands of individual actors who were never part of any collective deliberation. The UK National Cyber Security Centre warns that AI “lowers the barrier for novice cyber criminals, hackers-for-hire and hacktivists to carry out effective access and information gathering operations” (National Cyber Security Centre, “Key judgements,” par. 7). It also notes that “state and non-state, skilled and less skilled” cyber threat actors are already using AI in varying degrees (National Cyber Security Centre, “Key judgements,” par. 3). Anthropic’s 2025 threat intelligence report states the point even more directly: AI “lowers the barriers to sophisticated cybercrime,” allowing actors with “few technical skills” to conduct operations that previously required “years of training” (Anthropic 3). What was once a function of skill and coordination is now within reach of a single motivated participant.

For Bellaby's framework, this means that proportionality becomes harder to preserve when one participant can escalate from protest to damaging intrusion with little technical preparation. Discrimination becomes harder to preserve when AI-assisted reconnaissance can surface targets faster than any deliberative process can meaningfully vet them. Deliberation weakens when automated tools allow action to begin before consensus can form. The cumulative effect is that the conditions Bellaby's framework depends on become harder to maintain at the same time. AI therefore makes the weakness in Bellaby's framework more urgent: if ethical hacking depends on collective restraint, then a decentralized collective must have some way to preserve that restraint against individual participants acting in its name.

Bellaby's framework remains valuable because it gives political hacking a serious ethical structure rather than dismissing it as inherently criminal or accepting it as automatically justified protest. His standards of proportionality, discrimination, and deliberation identify the right questions: whether the hack responds to a serious harm, whether it targets the responsible actors, and whether it is shaped by public reasoning rather than personal impulse. As decentralized movements like Anonymous show, the harder problem is not deciding what counts as ethical hacking, but preserving those ethical limits once an operation begins. When anyone can act under the collective identity, a single participant can transform a broadly justified operation into one that appears reckless and unethical. Attribution problems and voluntary accountability make that weakness harder to solve, and AI raises the stakes by giving individual actors more capacity to escalate without collective approval.

For that reason, Bellaby's framework should be treated as necessary but incomplete. Political hacking may still be ethically defensible in rare cases of state failure or serious injustice, but only if the collective has some credible way to preserve proportionality, discrimination, and

deliberation throughout the operation itself, not merely judge them after the fact. Without that added layer of restraint, hacktivism risks losing the very moral distinction Bellaby wants to protect: the difference between political hacking as justified resistance and hacking as unaccountable harm.

Works Cited

- Anthropic. "Threat Intelligence Report: August 2025." *Anthropic*, Aug. 2025,
www-cdn.anthropic.com/b2a76c6f6992465c09a6f2fce282f6c0cea8c200.pdf.
- Bellaby, Ross W. "An Ethical Framework for Hacking Operations." *Ethical Theory and Moral Practice*, vol. 24, no. 1, Feb. 2021, pp. 231–55, doi:10.1007/s10677-021-10166-8.
- Loh, Wulf. "Anonymity, Fidelity to Law, and Digital Civil Disobedience." *Philosophy & Social Criticism*, vol. 49, no. 4, Mar. 2022, pp. 448–76, doi:10.1177/01914537211072886.
- National Cyber Security Centre. "The Near-Term Impact of AI on the Cyber Threat." *National Cyber Security Centre*, 24 Jan. 2024,
<https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat>.
- Rid, Thomas, and Ben Buchanan. "Attributing Cyber Attacks." *Journal of Strategic Studies*, vol. 38, no. 1–2, Dec. 2014, pp. 4–37, doi:10.1080/01402390.2014.977382.
- Scott, Tennille W., and O. Shawn Cupp. "Ethics of Hacktivism." *The Ethics of Future Warfare*, edited by Elizabeth Ditsch, Arthur D. Simons Center for Interagency Cooperation, 2018, pp. 143–148.