Limits and Possibilities for "Ethical AI" in Open Source: A Study of Deepfakes

David Gray Widder
dwidder@cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Dawn Nafus
dawn.nafus@intel.com
Intel Labs, Intel Corp.
Hillsboro, OR, USA

Laura Dabbish
dabbish@cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

James Herbsleb
jim.herbsleb@gmail.com
Carnegie Mellon University
Pittsburgh, PA, USA

ABSTRACT

Open source software communities are a significant site of AI development, but "Ethical AI" discourses largely focus on the problems that arise in software produced by private companies. Design, policy and tooling interventions to encourage "Ethical AI" based on studies in private companies risk being ill-suited for an open source context, which operates under radically different organizational structures, cultural norms, and incentives.

In this paper, we show that significant and understudied harms and possibilities originate from differing practices of transparency and accountability in the open source community. We conducted an interview study of an AI-enabled open source Deepfake project to understand how members of that community reason about the ethics of their work. We found that notions of the "Freedom 0" to use code without any restriction, alongside beliefs about technology neutrality and technological inevitability, were central to how community members framed their responsibilities, and the actions they believed were and were not available to them. We propose a continuum between harms resulting from how a system is implemented versus how it is used, and show how commitments to radical transparency in open source allow great ethical scrutiny for harms wrought by implementation bugs, but allow harms through (mis)use to proliferate, requiring a deeper toolbox for disincentivizing harmful use. We discuss how an assumption of control over downstream uses is often implicit in discourses of "Ethical AI", but outline alternative possibilities for action in cases such as open source where this assumption may not hold.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Social and professional topics** → **Socio-technical systems**; • **Security and privacy** → **Social aspects of security and privacy**.

KEYWORDS

deepfakes, ethics, open source, free software, agency, responsibility, interview

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533779

ACM Reference Format:
David Gray Widder, Dawn Nafus, Laura Dabbish, and James Herbsleb. 2022. Limits and Possibilities for "Ethical AI" in Open Source: A Study of Deepfakes. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3531146.3533779

1 INTRODUCTION AND RELATED WORK

Discourses of "Ethical AI" have largely focused on issues that arise in software produced by private companies. The drafters of the frequently cited "Montréal Declaration for a Responsible Development of Artificial Intelligence" [4] asked if we must "fight against the concentration of power and wealth in the hands of a small number of AI companies" in early deliberative discussions [5, 30]. However, an important perspective and site of AI practice is largely missing from "Ethical AI" discourse: Free and Open Source[1] developers creating AI software, who have unique limitations on and possibilities for ethical action. Open source AI development is significant: for example, two of the most popular AI libraries are open source: SciKit learn, and TensorFlow (after being open sourced by Google), along with myriad end-user AI projects. While harm does originate from a concentration of AI power in companies [85], we show that significant and understudied harms originate from differing practices of transparency and accountability in the open source community.

A 2019 systematic analysis of 84 "Ethical AI" guidelines [40] found that most guidelines are produced by private companies (22.6%) or governments (21.4%) often seeking to regulate AI from private companies. Abstract "Ethical AI" principles (*e.g.*, "transparency", "interpretability") are used with differing underlying meanings, and apparent convergence may be superficial [40, 47]. Systems may adhere to such principles while still being patently unethical [42], and convergence on principles risks obscuring political and normative disagreements [55], or focuses "Ethical AI" scrutiny on AI design rather than the business uses it enables [30]. Even critical discourse often focuses exclusively on the private sector: one study found that "principles alone cannot guarantee Ethical AI", but stated in their introduction: "AI is largely developed by the private sector" [55].

When design, policy and tooling interventions to encourage "Ethical AI" are built with private companies in mind, they risk being ill-suited for an open source context. For example, facing employee rebellion, Google decided to stop providing the US military with AI which could be used to improve drone strike targeting [82]. This decision was undoubtedly *politically* fraught, but enacting it was *procedurally* easy: the company exercised its legally available and

[1] *Open Source* eschews *Free* software's ideology; we use "open source" here. See Sec. 4.1.

enforceable right to not renew a contract. However, open source supply chains are messy: code is reused, and projects are copied and adapted (forked) [90], and it is difficult to track, constrain, or assign accountability for downstream uses. Conventional notions of accountability rely on stable entity to hold accountable, whereas open source membership can be unstable [59], and some even contribute anonymously [24].

Crucially, these structural challenges have cultural underpinnings. [20] The founders of the influential Free Software movement advocate for "Freedom 0"– the right of anyone to reuse code, for any purpose [72], encoded into legally binding licenses – and decry attempts to abridge this freedom even in service of other ethical ends [71]. Similarly, Transparency is often held as a near-universal principle in "Ethical AI" guidelines [40], but others reason how openness may not be universally desirable, giving autonomous weapons development as one example [15].

Studies to help AI practitioners improve fairness [22, 34], such as checklists to solve organizational challenges [49], are often based on the needs of AI practitioners in private companies, but some studies also focus on the needs of public sector [80] or academic institutions [61]. These results expose the role of organizational structures in AI Ethics practice, structures which look very different in open source. On the other hand, incentives in private organizations can hinder "Ethical AI", where developers work in "an environment which constantly pressures them to cut costs, increase profit and deliver higher quality [systems]" [79], and "face pressure from management to make decisions that prioritize company interests"[51, 55], and companies compete in a wider market structure which can hinder "Ethical AI" work [53]. Alongside the possible challenges for "Ethical AI" in open source we discuss in our paper, we also see a cause for optimism: unconstrained by these forces, experimentation may be more possible in open source communities to offer new ideas to solve ethical challenges unsolved in company contexts, or provide space to challenge assumptions made in private companies' "Ethical AI" endeavors.

To begin reconciling conflicts between norms in open source communities and prevailing assumptions in "Ethical AI" discourse, we ask: **How do members of an AI-enabled open source Deepfake project reason about the ethics of their work?** To answer this, we conduct an interview study in an open source community which builds software to create "Deepfakes": videos which replace the likeness of one person with another [43]. The community celebrates artistic and educational uses they see as ethical, and explicitly takes a position against and actions to discourage uses they believe are unethical, such as non-consensual or child pornography and fake news. In our study, we uncover normative, structural, and technical barriers to the community achieving their stated ethical views, and situate these barriers within the dominant private-company-focused "Ethical AI" discourse and political tensions in the open source and wider tech worker communities. In the additional appendix, we outline ideas that open source communities and platforms may want to experiment with, which researchers may also be interested in evaluating and studying further.

2 METHODS AND SETTING

2.1 Setting

We set our study in an open source Deepfake creation tool, an AI technology with contested ethical issues [43], positioning it as an extreme case study [89] where ethical reasoning and its situated relationship to other cultural frames may be especially apparent. A 2019 study found that 96% of online Deepfakes are non-consensual pornography, 99% of which depict women celebrities [6]. Scholars write that political Deepfakes operates similarly to non-consensual Deepfake pornography to silence critical speech, and that victims of the latter experience anxiety, illness and job loss [50]. Other scholars explore how Deepfake distribution enforces gendered disparities in visual information [81], and find that more attention in public discourse is given to viewers of Deeepfake disinformation than do the women depicted in Deepfake porn [29]. One study analyzed Reddit and GitHub posts and found a tension between between moderation practices and open source ethos, recommending future work beyond identifying or regulating Deepfakes to understanding their antecedent code and programmers which enable their creation. [86]. We do not seek to define or evaluate ethical behavior, which others studying AI practitioner's views on ethics (*i.e.*[61]) recognize as an entire branch of Philosophy, with divergent proposed approaches in AI [10, 13, 30, 57]. Instead, we examine "how AI practitioners understand the ethical landscape and their own role within it" [61], including "procedures, decisions [... and resulting] related responsibilities" [61], and examine how their perspectives do or do not fit with prevailing AI ethics discourse.

The first widely-available face swapping algorithm was posted by an anonymous user in a Deepfake-focused sub-Reddit [6, 43, 86], which has since been banned for violating the site's more recent "policy against involuntary pornography". This algorithm became the basis of many open source projects; we approached the project of our study because of its unique willingness to engage in questions of ethics as indicated by its public ethical stand. The project's original leader copied this algorithm from Reddit to a repository on the social coding platform GitHub [23], which new leaders use to track code changes and bugs, and host usage instructions and a contributor guide. Current leaders rewrote the codebase and applied a GNU General Public License (for implications of this, see Sec. 3.1.1). The GitHub project page prominently features a statement written by the project's leaders to explain the benefits of releasing the software publicly, such as enabling AI learning, political commentary, and artistic uses, while acknowledging and claiming a refusal to support non-consensual, inappropriate, illegal, unethical, or questionable uses. The GitHub project page directs support requests to two other platforms: a Discord chat server and a self-hosted online message board. On all platforms, there is an expressed "Safe For Work" policy, for example, one is posted in the "Welcome" section of the Discord chat server, which states that even discussing NSFW content will result in an immediate ban without further warning. These platforms provide space for the 500+ users who are often online at once to seek and provide technical support, share Deepfakes they have created, and discuss broader Deepfakes and AI issues. The leaders are informally designated, often being invited to join private channels and given administrative privileges by existing leadership after contributing to the project codebase, or by creating

high quality Deepfake content. These leaders use these channels to discuss development and moderation decisions, which they have broad discretion to make independently. This project is not corporate affiliated, but accepts donations. Users often used humorous display names, but established users often knew each other's real names. The first author observed a generally collaborative and polite tone in these venues.

2.2 Recruitment, Participants, and Data Analysis

The first author approached project leaders who gave permission to recruit in their community and collaboratively crafted a recruitment message which a leader shared in the project's chat server, resulting in eleven completed interviews. All self-identified as male, and were mostly from the United States and Europe, resembling open source generally, but the modal age range was 35-44, somewhat older than open source generally [28]. For confidentiality, we do not discuss individual demographics. Participants had a median of 7 years of programming experience, 2 years of AI experience, and 2 years working with the project in roles ranging from developing and testing the project code, supporting users, content moderation in communication channels, and both hobbyist and professional users of the project. The first author conducted semi-structured interviews in a one-on-one setting due to the possibly sensitive nature of topics discussed [84]. Most interviews were conducted via a teleconferencing call and lasted 30-93 minutes, with most lasting about an hour. In two cases, chat interviews (*i.e.* [74]) were used for accessibility reasons.

We adopt an *interpretivist* epistemological paradigm [46]: the framings presented below emerge from the intersubjectivity between researcher and participant, and cultural frames they do and do not share. We also observed chat room discussion and work interactions on GitHub, but we acknowledge that self-reports from our primary interview method may hold limited value in explaining behavior and attitudes in actual context [38], and caution that there are meaningful differences between open source communities that limit the ability to generalize these findings to the exceptionally organizationally and politically diverse landscape that is open source. We note that the male dominance in this community and Deepfake production communities generally contrast sharply with the vast majority of online Deepfakes which non-consensually depict women in pornography [6], and past work which we discuss in Section 2.1 has discussed the gendered politics of Deepfakes, and future papers using feminist analytical frames could unpack gender dynamics of how exclusion plays a role in the choices that open source communities see as available to them.

Data was analyzed in an iterative process including a descriptive memo after each interview, and a running analytic memo as a reflexive history of the first author's understanding of emergent themes [75], and weekly discussions among the research team to discuss commonalities and contrasts between interviews. After data collection, all interviews were transcribed, and then the first author examined possible relationships between themes in this analytic memo, iteratively going back to the data to test out these possible structures, before settling on an inductive hierarchical coding frame [48, 54, 76]. This was then used to code the entire dataset. During this coding process, our understanding of the data deepened and new codes arose to capture new themes or provide greater specificity, in which case an open card sort was used [64] to identify sub-codes, after which the dataset was re-coded.

3 FINDINGS

3.1 Responses to Ethical Issues

Participant's perceptions about what they could and couldn't do about Deepfake misuse was shaped by open source licensing, discourses about progress and the neutrality of tools, and by setting community norms of acceptable use.

3.1.1 Open Source Licensing as a Frame for Ethics. We saw that the open source license of this project is highly relevant to participants' ways of understanding their responsibilities, and therefore their responses to the problem. It is both a legal set of constraints that sets out what developers can and cannot do to prevent uses they view as unethical, and a normative one that frames broader cultural values beyond what the license requires (see also [20, 41]). Leaders lamented that, as they saw things, the open source status of their project (a choice they made) prohibits them from controlling downstream uses. A leader remarked: *"We've got very limited control. [...] We can't prevent people from getting access to a software using it. [...] Part of being open source Free Software is that you are free to use it. There are no restrictions on it. And we can't do anything about that."*

Even if the leaders wanted to choose a more restrictive license for their project, the leaders' prior choice of a GPL license led contributors to view applying a more restrictive license as impossible at this stage: *"Anything that touches GPL code becomes GPL code, right? There is no takesies-backsies. There is no reversal."* However, the issue is not just about the GPL as a legal requirement, but the norm that it sets. When a project moderator was asked if anyone had considered rethinking the project's open source status to control how it is used, he said that this would *"kill the project"*, and that this would mean that the project gets less *"free help"* and ideas. Another contributor stated that would require a lot of labor to do in a *"moral"* way: *"rewrite the whole thing from scratch to make it closed source."* Community members did not not seriously consider alternatives to open source licenses.

Participants also used the open source license as a reference point in reasoning about incorporating technical restrictions on problematic use. Leaders discussed an image recognition based content filter that would prohibit the software being used to create pornographic content, or embedding a visible or encoded watermark identifying the video as a Deepfake to enable people to distinguish between doctored and real footage. However, many participants believed there would be *"no point"* putting in restrictions because the project's open source status means such safeguards could be easily removed: *"I cannot stop people [from] using my software for stuff which I don't agree with [... open source's] positive is also it's negative: [...] anyone can read all the source code and then can change any of the source code they want [...] whilst you can build stuff in to maybe stop your software being used in the way you want, someone [can] just rip it out again."* Other participants believed *"forcing"* such restrictions would require them to *"actively invade our user systems"*, reflecting not only a practical but moral aversion.

"Forking" projects—copying the code into a new repository and working on it anew—is frequent in open source [90], which has the effect of distributing and decentralizing control [20]. This led another leader to believe that forking would lead to an additional, separate community without the ethical guidelines and content moderation they use: *"Let's say I built a load of limitations into my software [...] and anyone who used it, uh, would fall afoul of those filters. Well, what should happen is that the code would be forked and then everyone would start using the fork [...] And what effect does that have? It takes people to a version of the software, which doesn't have the ethical guidelines and doesn't have the moderation in place to make sure people aren't using for that. So you're kind of shooting yourself in the foot."*

Another participant recalled when GitHub removed a project used by music pirates [19], leading to broad proliferation of that project's code (*i.e.*, Streisand effect [37]) and expressed that restricting access would thus backfire. Another also believed this: *"If it was shut down, if the code would be deleted from GitHub, everyone would have it still on their computer and it would be easily find-able on the dark web."*.

Decentralized control in open source also makes some technical approaches to preventing harm more difficult, as one participant explained: *"Some of these server-based [deepfake] apps [...] actually have filters [for] nude pictures. [...] That's a different kind of setup because [...] they're taking photos that people are uploading then processing them on a server then spitting them back down to the user. So because of the centralized control [...] they could implement filters. I don't know that it could be practically implemented in an open source project that isn't server-based."*

Finally, the transparency to examine source code provided for by the open source license was seen as an important resource for overcoming some types of harms. For example, a participant explained someone had embedded malware in a closed source app made using the original face swapping algorithm: *"he started putting a crypto miner in the program. [...] any closed source application like that in a relatively niche area has the potential for someone to put some sort of illicit material in there"*

3.1.2 "This genie's out of the bottle": Technological Inevitability. Many participants believed that because the original Deepfake algorithm is widely circulated, further development of Deepfake technology is inevitable, arguing that halting their own development work or other restrictions would only *"delay"* development, but would ultimately be ineffective. One stated: *"if our project shut down today, deleted everything, there are other ways of [creating Deepfakes]. I mean, there are several other ways, uh, and you see them pop up, like I've [seen another app] and [another open source project], there's another piece of software and there are others."* When discussing that their project had likely been used to attempt to influence an election, one project leader stated: *"if it weren't for [our project], they would have [another app...] It's not like the amount of work that it takes to make a face swap is far less than finding [our project] or one of its competitors"*. The same participant extended the alternatives idea from alternative projects to alternative individual contributors, referencing his involvement: *"In the end of the day we knew that that sort of thing was going to come about whether or not I participated in [the project]"*.

Some laws now criminalize non-consensual pornographic use of Deepfake technology [39]. Some participants viewed laws criminalizing the use of Deepfakes as naive given this inevitability, one saying those intent on unethical uses would not follow regulations anyway: *"Heavy-handed regulating is just going to hamstring us because there are countries and actors out there who just will do it [create Deepfake software] anyway, right? [...] If history has shown us anything, that if you ban something, it just goes underground"* Another invoked a genie metaphor to argue for the irreversibly of technical progress and express distaste for regulatory action: *"I also don't believe in like, just banning something because it could be dangerous. It's just, first of all, it's not going to work. You know, this genie's out of the bottle."*.

Historian of technology Arnold Pacey framed the technological imperative which fuels this feeling of inevitability as "the lure of always pushing toward the greatest feat of technical performance or complexity which is currently available" [62], and mathematician John von Neumann said that "technological possibilities are irresistible to man" [56]. Our participants appear to embrace this alluring inevitability, one participant referencing futurist Ray Kurzweil and then stating *"There's nothing that can be done to stop the steam engine that is progress. And technology, it's only getting better, faster"*.

Philosopher Daniel Chandler argues that surrendering to the the technological imperative "implies a suspension of ethical judgement or social control: individuals and society are seen as serving the requirements of a technological system which shapes their purposes", and that it is possible to abandon even "large, complex, interconnected and interdependent" technological systems, "given the political will" [17]. We see that our participants view their own role in developing Deepfake software as insignificant in the context of the wider progress of mutually interchangeable alternatives. They point to the proliferation of the original face swapping scripts before their specific project, and the broader idea of Deepfakes, as evidence. In a similar vein to the debate on nuclear proliferation [69], some participants framed these other parties as "competitors", and developing this technology as a race, thus making this needed widespread "political will" feel impossible. Implicitly, participants point out that to halt it all together, the political will to do so must be held by many uncoordinated open source, private company, and state actor developers of Deepfake software.

3.1.3 "If I painted something offensive, you can't blame the paint manufacturer": Just a Tool? Some participants stated that they view the project as a tool, and that the ethics of any particular use case is solely up to the user, in line with views expressed by academic, public and private sector AI practitioners [61]. One contributor stated: *"You can't really blame the project cause it's like blaming the people that make the paint and the canvas [...] You can't blame them directly by no means."*. This participant then localized this sentiment to their project specifically by comparing it to the image editing tool Photoshop: *"I mean we provide the tools, but then again, I mean, would you blame Photoshop if someone just put someone else's face on another's body? I mean, no! That's ridiculous."*. Others also employed the Photoshop comparison (which has also been discussed in past research [86]), stating while they believed Deepfaking has a greater ability to harm, the use of the technology is up to the conscience of the user: *"Face swapping is basically a more sophisticated application*

of, for example, using Photoshop to enhance the figure of a model. I think obviously it's more powerful and it has a greater potential to harm people, but I think the use of the technology has to be left to the individual conscience of the user". Others compared the project to to recent uses of long-criminalized psychedelics to treat depression, and cannabis to treat other medical issues, suggesting that it would be bad to *"hamstring a wonderful technology on the risk that a couple of bad actors will do something [bad]"'*

One of the project's posted statements explicitly states that the project can be used for "good" or bad, a property it claims is common of any technology , which alongside views expressed above, reveals an *instrumentalist* view: while the way a technology is used may have moral implications, the "technology [itself] is neutral, subservient to our beliefs and desires; it does not significantly constrain much less determine them" [70]. However, as we will see in Section 3.3.3, some participants acknowledge that project's design can influence how it is used. Another participant agreed that changing the project's design could make certain uses less likely, even if not impossible, by implementing technological restrictions into the code: *"For people that [want to make problematic pornography] they're not very into [...] how it works. They just want the end result. [...] Right now you have to do quite a bit of manual stuff and you have to set up the whole environment..."* Thus, he suggests that technological restrictions designed into the project *"could be a future idea that would stop a lot of people already"* from using the software unethically, except for the *"very good programmers [who] will be able to take that [restriction] out"*. This participant reached a conclusion similar to many before [36, 44, 45, 70] which we discuss in Section 4.2: *the design of tools make certain uses more or less likely, by requiring time and skilled labor to circumvent restrictions.* As we saw in Section 3.1.1, project leaders decided against restrictions for fear of their easy removal, but also worried that they may lead to splinter communities without the ethical norms we will now discuss.

3.1.4 Setting and Enforcing Counter-Norms by Denying Support. We saw in Section 3.1.1 that open source licenses shape views about developers' possibilities and responsibility for limiting downstream harm by presenting the right to use software for any purpose as paramount, but the project's leaders sought to set countervailing cultural norms to actively discourage uses they believe are unethical, without preventing such uses completely. There is a long history of open source communities setting norms outside those laid out by licensing, a process that Free Software anthropologist Chris Kelty describes as a "punt to culture" [41]: developers turning to persuasion, rather than strict, punitive control via legal or technical means.

The tactic of setting and enforcing counter norms is most clear in a public statement intentionally displayed as a *"very public policy"*, which states that they intend their project exclusively for "ethical uses", and that it is not for creating "inappropriate" content. One developer for the project reflected that this is difficult to enforce: *"One of the points in our [public statement] is that [the project] is not for changing faces without consent or with the intent of hiding its use [...] Again, we can't force our users to do anything."*. Enforcement, appears less important than articulating what does and does not count as harmful use in the eyes of the project. This has the effect of building consensus, which, in a distributed environment where projects can fork at any time, can be powerful. This tactic is also visible in the argument seen in the previous section that technological restrictions would make ethically undesirable uses *harder*, not as much a literal strict control as discouraging unethical use.

Leaders often expressed the view that denying valuable technical support [63, 83], to those attempting to create Deepfake porn is their only way to *discourage* such uses, absent being able to outright *prohibit* them given their understanding of the legal dictates of open source discussed in Section 3.1.4 and 3.1.1. This is shown by the quote: *"So there's not a lot actively I can do. [...] But what I can do is discourage it and not [...] offer advice, and actively block people looking for that advice within forums and domains that I have control over."*. Project leaders recounted when they have banned people for soliciting help to create Deepfakes that contravene their rules, often after discussing the offending case privately amongst other leaders first. Another leader stated that refusing support is the "best" means of control they have: *"Best we can do is say, we refuse to support you"*, going on to say *"if people are using it for that sort of thing, they're not going to tell us"*. Others framed this in terms of choices about their own labor, which fits squarely with open source notions of freedom: *"I don't need to teach anybody or learn how to put Scarlet Johansen's face on, you know, insert porn star here"*. Here, withholding of support became a matter of maintaining community, both in terms of who participates, what activities are acceptable, and how people choose to spend their time, which is not seen as in conflict with open source norms *per se*.

Combined, these efforts are having clear effects. Users of the software echoed the sentiment that the developers of the software are largely doing all they can to prevent misuse: *"I think there that they're probably doing all they can [...] it's not like they're going to be able to build like a detector or something for how the software is being used."*. The effect of these norms requires individual community members to take them as seriously as the users we spoke to did. Because these additional norms are not strict rules (anyone *can* use it for any purpose, per their GPL license), some weigh them against what they see as a higher purpose: the foundational norm of producing open source code. One project leader reported sometimes learning of pornographic uses of his software from crash logs, but reported overlooking this in favor of improving the software using these helpful logs: *"I try not to read what those are because they're not important for what I'm doing, but you could argue that I should ban people as soon as I see [them]. From my point of view, I want to make the software better. So the crash report is useful for me. And as I said, I can't stop people using it for reasons I don't agree with, but I can discourage you."*.

3.2 Motivations for Ethical Action

Participants expressed intrinsic motivators for wanting to prevent harm, namely commitment to their own ethical lines and extrinsic reputational costs.

3.2.1 Ethical Lines: Consent, Family, Law, and Professional Standards. One leader described the creation of the public statement expressing the project-wide norms of acceptable use as arising from a kind of spontaneous agreement: *"We just all happened to be in the same place"*. However, participants explained how they arrived

at this norm in a variety of ways: a commitment to consent and concern for the harm caused when it is violated, as well as a commitment to familial norms and professional and legal standards. Studies examining the motivation of open source developers on technical matters identify similar intrinsic or altruistic motivating commitments [8, 32].

Many participants demonstrated reverence towards the concept of consent. One participant spoke about how it is wrong to non consensually use someone's identity to sell products, saying *"you can't steal a celebrity's likeness to sell a product, right?"*. Another professional Deepfake creator created a Deepfake of a deceased person on the request of their relatives, but expressed ethical concern about whether this respects the deceased person's consent. One participant discussed how *"consensual pornography is completely up to the people involved"* and a project leader echoed this:*"I don't have an issue with porn."*, but then explained their own support for the blanket ban on asking for help creating porn because of practical and moral complications in ascertaining consent: *"It might be their wife and they have some weird [Deepfake] fetish. Okay. That's their thing. [but] It might be the neighbor's 12 year old girl that they got the hots for and have been videoing from a distance. No, [...] I'm not going to take the time to sit down and [say] Oh, maybe there's a gray area."*. However, others believed that it is ethically permissible to create porn of someone else without their consent, because they believe sharing it is where most of the harm may lie: *"I think that's, it's okay to enjoy whatever you want, as long as you don't hurt other people with it, [...] obviously posting it online for other people to see and potentially for the person you don't have consent for, to find out that that will have a negative effect on them."*

Others tied their personal sense of morality to how members of their family may react to certain uses. One participant initially said *"I really don't know how to define what's right and wrong"* but then proposed a standard by asking *"would I show my mom?"*. Another participant stepped up a generation to suggest a litmus test to catch possible fake news: *"if you tell your grandma about it and you fooled her, and she thinks it's real, but it's a fake and it's saying something negative about someone else that's, that's not kosher"*.

Finally, others invoked professional and legal standards when discussing their personal sense of ethics. One participant who operated a Deepfake based marketing firm discussed a "very clear" line for his firm, informed by his experience as a photojournalist: *"We don't cross the line. [...] We follow things like [...] various journalism association standards and normal things you would follow if you're a Washington DC political correspondent"*. Another professional Deepfake creator declines pornographic Deepfake requests by explaining to prospective clients that such uses may be impermissible under law.

3.2.2 Reputation. Past research shows reputation motivates open source contributors and influences their behavior: open source contributors actively promote their contributions to gain status [23], reputation is important one's contributions being accepted [35], and that job candidates and employers see contributions as indicators of technical skill [52]. Here, reputation motivated ethical action at the personal level for hobbyists to label Deepfakes as such and for professionals to attract business; in the project we study to protect itself from censorship and differentiate itself from competing projects with perceived less ethical behavior; and the wider professional Deepfake community to escape the stigma of Deepfakes.

At the *personal* level, hobbyists strive for realism to show off that they are creating realistic Deepfakes, which calls attention to its fictional nature: *"if I could ever achieve [...] undetectable realism, then obviously I was gonna make a big [...] hoo-ha about it!"*. Another explained why most Deepfakes are labeled as such, reducing the risk of fooling people in his view: *"Truly cutting edge [Deepfakes] are presented in a context that highlights the fake rather than disguises it, which is no surprise as the poor sod who's worked on it would naturally want to draw focus on their effort."* Similarly, professional Deepfake creators reported creating high quality fan-art Deepfake content to post online to demonstrate their skills, get exposure, and get business. These people advertise a Deepfake explicitly as such for reputational gain, and these participants believe this mitigates risks of fooling viewers.

At the *project* level, leaders have gone to great lengths to protect the reputation of their project, because it had been previously delisted from Google results, put behind a login wall on GitHub, and had members banned from their Discord because of associations with non-consensual pornography in the media. One leader reports that the project's public statements were in response to the project being delisted and blocked. He also worked with GitHub to remove porn and porn-related images from GitHub issue threads created before he led the project, and adopted a contributor Code of Conduct to defend the reputation of his project and as a condition for GitHub to remove the login wall from their repository. Another leader explains that *"We don't want [the project] to be identified as hostile [...] We want people to be able to find us and find the software without having to face a deluge of nonconsensual pornography"*. A user of the software echoes this, saying the public statements are a *"very good"* idea because then *"the media doesn't think that there's a group of programmers just trying to create blackmail software. Then it might've been shut down by GitHub."*. We see that the leaders of the project engage in activities to limit unethical uses of their software partly in response to enforcement actions by the platforms they depend on, implications of which we discuss in Section A.

The leaders reported a feeling of unfairness, pointing out that another Deepfake project's Github page links directly to a porn website and its forums to provide technical support, yet it apparently has not faced the same restrictions or had to do the same work to maintain a clear reputation. At the same time, when one leader is asked how he'd feel if his project was used for something he disagreed with, he replied *"I don't think I'd feel particularly bad about it because I'm not naïve [but if something went viral with his project's name attached] that would bother me, because that would be an association with my product"*.

Finally, at the *professional community* level participants who were members of the professional Deepfake community expressed an interest in protecting the ethical legitimacy of Deepfaking as a practice. One participant who is part of a small community of highly-skilled professional Deepfakers said *"[it is] frustrating because everyone that I know that's [creating Deepfakes] is doing it for the creative possibilities, to explore the ethical uses of [Deepfakes]. And it's like, you know, it's an uphill battle because of the sensationalism, um, about Deepfakes"*, further describing the competing open

source project which promotes the creation of non-consensual porn as *"unprofessional"* Another participant explained that this negative reputation is *"a large part why most of those within the community [...] tend to be rather hostile towards those who show up asking for tips on how to create [pornographic content]"*. One participant explained that they have attempted to rebrand: *"a lot of us 'Deepfake' artists have come around to preferring the term 'synthetic media' [...] leaving the stigma of "Deepfake" behind."*. A casual user of the software expressed empathy with professional content creators: *"It's an association no one wants, to have the effort put into creative works using the tech marred by the association with these less than respectable use-cases is certainly no fun for content creators"*.

3.3 The (in)Accessibility of Deepfake Realism

We found that Deepfake realism is prized, and some suggested that more people should have access to this artistic tool, while others argued that difficulty achieving realism mitigates societal issues.

3.3.1 Deepfakes for the "Everyman". Participants celebrated that the ability to create Deepfakes is now broadly accessible to everyone, not just to those in academia or in companies with special training and technology. One participant stated: *"There's something quite thrilling about the everyman (sic) having access to the tools to create results that depending on hardware could be on par with what industry professionals might cook up"*. We note that the gendered term "everyman" betrays something participants did not address directly: that these are tools made and used largely by men. Some did, however, recognize the harms to women associated with misuse. We discuss this briefly in 2.2 and point to literature discussing intersections between gender and Deepfakes in 2.1.

Nevertheless, widespread access was seen as a self-evident good: *"Machine Learning is an incredibly complex process which generally is the remit of academics. And so my drive for developing [this project] is to basically take this kind of impenetrable area of computer science and try and make it as accessible as possible for people."* Echoing this sentiment, a professional Deepfake creator speculates that the output possible from a competing open source Deepfake project is equal if not superior to the work that leading visual effects firms are capable of: *"I don't think there's another program that you can get open source that can do what [open source project] does. I imagine like maybe Disney and ILM [a visual effects company] have home-built tools that can compete with it, but I honestly don't think [they do]."* This sentiment is crystallized in public statements on the project, which portray AI as exclusive knowledge, documented in arcane research venues, but that their project opened participation to all.

This impulse to "democratize" access to an inaccessible technology by wresting it from the hands of an exclusive few for the benefit of common folk is an ethical ideal which sparked the Free Software movement [20, 71]. This is a different notion of democratization than those seen elsewhere: a minority of "Ethical AI" guidelines from companies and governments reference political ideals such as open dialogue, broad participation and wider principles of democracy [40], and private companies are increasingly co-opting similar political language when marketing their AI endeavors [16]. Interpreted in the context of this wider political landscape, some of our participants accept the possibility that their software is used unethically to prioritize an ethically charged commitment to democratization.

3.3.2 Inoculation through Proliferation: More Deepfakes as Remedy. Some participants argued that the antidote to ethical issues stemming from Deepfakes, such as fake news videos or defamatory porn, is increasing skepticism and distrust of videos which will be brought on by the deliberate and increasing proliferation of Deepfakes into the popular consciousness, whereas keeping them *"locked away would do more harm than good."* This sentiment is expressed by the leaders of the project, one saying: *"One good reason to promote the use of Deepfakes in satire and in various other areas is inoculation: teaching people not to just blindly believe what they see."* By analogy to Photoshop, one participant explores a world in which Deepfakes are not widely known or accessible: *"Imagine a fictional world in which Photoshop as we know it today is something only accessible to a select few industry experts with a budget of hundreds of thousands if not millions. Due to the far reduced exposure that the everyman might have to the works that can be created with Photoshop they would be far less liable to question a doctored photo when seeing one."*

In this way, participants argue that *"ubiquitous"* proliferation of Deepfakes becomes the cure to the harms this proliferation may bring, by "inoculating" people: making them not trust videos they come across without further verification.

3.3.3 Low Accessibility and Realism is a Safeguard. The previous two discourses saw access to Deepfake realism a greater good or even a way to prevent harm, but disagreed: some argued that extreme ethical concern is unwarranted because the high effort needed to make realistic Deepfakes prevents some bad actors from using it for ill, and that unrealistic Deepfakes unlikely to fool people. For example, a minor contributor to the project speculated that: *"They're not making it more accessible, I think on purpose to weed out the people that don't know a lot about technology and just want to do it for bad intentions."*Another participant who Deepfaked President Biden with dubious realism stated that he thinks those with political agendas are unlikely to expend the effort required to make realistic Deepfakes: *"I put Biden as the Trololol guy [an internet meme] and you can look and it's not great, but it's funny, you know, and that's about, yeah, I don't think anybody with a political agenda of some form is going to put much more effort than I did into it. So you're going to be able to tell [it is] fake. So it's not like it's going to change the direction of a country or something like that."* One of the project's leaders stated that though he wishes people would explicitly mark Deepfaked videos as such, he thinks they are implicitly marked because they are often low quality: *"I feel like it is clearly marked even if they don't put it in the tags, because Deepfake quality is not really there."* The project's leaders *are* focused on improving the quality and realism of the results, however, so any ethical benefit of having Deepfakes marked by their low realism may not persist.

Most considered professional work to be quite distinct from home-made Deepfakes. A moderator of the project referenced the movie Avatar, lauded for its visual effects [1], to explain that convincing fakes have long been possible with a large production team, convincing homemade fakes will be rare. Professional Deepfake creators describe those dedicated to highest quality as a small community analogous to the early days of long exposure photography: *"you have to be a pristine technician in handling all the parameters to*

set up your camera and everything". Similarly, a user of the project said he's never seen a Deepfake that he thinks could fool people, but high cost will prevent this for the foreseeable future: *"convincing higher-resolution models require exponentially more high speed video memory. As it stands this is not cheap at all, and won't get cheaper for some time still."* Here, participants are assuming that technical prowess or access to expensive hardware aligns with ethical scruples: people who can overcome technical hurdles to create convincing Deepfakes are less likely to create ethically problematic Deepfakes.

While the sentiment *"I've never seen it to be done realistic enough to pose any sort of ethical issue"* appeared widespread, one participant expressed fear about the project enabling widespread, indiscernible Deepfakes: *"If [this project] is that accessible and that, because computers will get better, everyone can do it on their phone and in a bunch of years. It'll be scary if video evidence would never be trustable anymore."*

4 DISCUSSION

4.1 Helpless to Challenge Freedom 0? Limits and Possibilities for Developer Agency

Many participants felt unable to control downstream uses of their software, given the dictates of Freedom 0 – a core principle of Free and Open Source Software which demands that users should be allowed to use the software for any purpose, and is a primary way open source "democratizes" [20]. Throughout the research, we saw that Freedom 0 was treated as an unquestioned default norm more so than an accidental effect resulting from a mere choice of license. Freedom 0 is so fundamental that it is even encoded into the platforms that projects depend on. For example, the code sharing site GitHub's license picker only points to licenses where Freedom 0 is protected justifying this with the pithy statement "An open source license protects contributors and users. Businesses and savvy developers won't touch a project without this protection."[2] The strength of this norm meant that at times participants expressed either misconceptions about the implications of their license choice, or their ability to have chosen a different license, or mistrust that people would abide by alternative license terms. For example, the same leader who lamented that they cannot control what people use it for was involved in choosing the license which inscribes this relinquishing of control. This project chose a General Public license (GPL), a "Copyleft" license where publicly-available derivatives or subsequent versions of the software must be distributed under the same freely-released terms. The choice is indeed effectively irreversible without the consent of the many anonymous past contributors, but this leader did not articulate his own agency in the first act of choosing a license. Similarly, we also see Freedom 0 at work in the tendency to view technical controls in literalist terms, and therefore to find them ineffectual rather than norm-setting.

The norm of Freedom 0 underscores and elaborates other discourses like Tool Neutrality and Technological Inevitability, which also frame designers and developers as lacking agency. These discourses are also common in proprietary contexts, but there, the ability to choose among or create bespoke closed source licenses is more visible and common because there are other concerns (such as liability, financial obligations, or regulatory requirements) that make the need to limit uses (such as, to paying customers) more common and apparent. Where action is taking place in OSS, it is happening via other discourses, such as setting counter norms and making choices about where one's unpaid labor goes.

The project we studied was not prepared to question Freedom 0. However, Freedom 0 is situated in a changing field of claims and counterclaims about software ethics. This field has a long history, including the contentious term "open source" itself, which represents a change from the early days of critiquing of business practices that restricted access to source code [20], towards the promotion of "open source ideas on 'pragmatic, business-case grounds' " [77]. Just as practices and licenses changed in this previous shift, it is possible that projects committed to Freedom 0 may be forced to respond to newer changes. For example, the Ethical Source movement, part of a broader reckoning inside and outside tech companies, was founded to participate in "giving [developers] the freedom and agency to ensure that [their] work is being used for social good and in service of human rights"[3]. This centers the *developer's* freedom to choose how the product of their labor is used, away from the *user's* freedom to use the software for "any purpose", with the goal of using licenses to foster that "make it easy for the user to do the right thing" [41]. This recentering does appear to call for a rethink of Freedom absolutism. Other developments have similarly recentered the importance of developer labor rights. The *Tech Won't Build It* movement "holds that workers developing AI/ML should have a say in how such technologies are deployed" [73], and the Tech Workers Coalition advocates (among other things) for workers to have a say in how the products of their labor serve "people, communities, and the environment rather than solely [...] profit" [3], aligning with the Ethical Source movement's framing of freedom. Whether commitments to Freedom 0 change in light of these broader changes, is a key question for the future.

4.2 Transparency and Accountability for Implementation vs Use Based Harms

AI systems can cause harm in multiple ways, and locating the causes of each harm on a continuum between *implementation* and *use* may be conceptually useful in debates on how to mitigate them. We define *Implementation harms* as those arising through code, algorithm or data problems that can be fixed without changing the intent, or use, of the software, for example through the use of "de-biasing" techniques to reduce bias in algorithms or training data [9, 18, 78]. On the other hand, we define *use based harms* as arising from a use which may *itself* be harmful, that no amount of technological fixes, implementation improvements, or more or better code will alleviate the ethical concerns with the software. Some make this point through satire, showing how dilligently followed "ethical" implementation fixes do not alleviate the patently harmful use of mulching elderly people [42]. Others find that corporate-backed AI values statements focus more on AI design decisions (implementation) than questioning the business *uses* which AI enables [30].

Our open source case demonstrates how harms originating from each end of the use-implementation continuum are differentially

[2]See: https://choosealicense.com; some licenses require the free sharing of resulting derivative code, which companies may desire to keep proprietary.

[3]See: https://ethicalsource.dev

affected by the limitations of transparency [11]. Free and Open Source software offers accountability through individual traceability to specific lines of code. Grodzinsky *et al* wrote in 2003 that the "many hands" problem (*i.e.*, collective responsibility [12, 27]) in software development can lead to "harm and risks for which no one is answerable and about which nothing is done" [58], but argued that open source enables individual-level accountability because "if a developer were to write irresponsible code, others contributing to the open source software would be unlikely to accept it. [...] Parts of code can be ascribed to various developers, and their peers hold them accountable for their contributions" [31]. This traceability indeed helps identify and rectify *implementation* harms that occur through code quality issues, as exemplified by our one participant's reference to a surreptitious cryptominer in an alternative closed-source Deepfake app, and the transparency that open source facilitates allows scrutiny which can help illuminate and mitigate unfairness in classification or prediction systems, arguably harder to accomplish when the model and data is proprietary [14, 66].

However, our findings show that Open Source has less power to support accountability for *use* based harms, because harm can be wrought not only from parts of code which may malfunction or be ethically inadequate in some way, but from the whole software package operating as its creators intend, but for a harmful use they did not intend. Notions of transparency in open source combine access for scrutiny purposes (referred to as Freedom 1 in the Free Software community [72]) with unconstrained use, circulation, and modification (codified in Freedoms 0, 2, and 3 [72]), a combination which allows *use*-based harms to proliferate. In our example of Deepfakes, open source's transparency and unconstrained circulation can help such harms proliferate, allowing unscrupulous users to learn the relevant techniques and achieve their goals without the "friction" of rebuilding code from scratch. In short: open source's commitment to transparency of implementation allows strong accountability for implementation-based harms, whereas the same commitment to transparency allows use-based harms to proliferate, and absent a matching commitment to transparency of use which would make such harms visible, leaves it powerless to support similar accountability of use.

The risk of this openness aiding the proliferation of potientially harmful technology such as superhuman AI [15], and claims that open source contributors are unacceptably expected to abrogate control over the ethical impact of their creations [87] have been explored before, and we unpack how open source norms lead some contributors to accept similar risks. Others suggest that market logic will operate in open source development to prevent harm because "'good guy' AIs" will "out compete the malicious and incompetent" [33], echoing the trust that some AI practitioners place in market logic to diminish less trustworthy AI [61], but we instead find that this competition lead some participants to view ethically mitigating practices as futile (see Section 3.1.2).

Of course, implementation is not always cleanly divorced from use: the designers' intent, the affordances they implement, and the influence these affordances have on users change the likelihood of unintended use. For example, our participants disagreed whether the Deepfake software was a "Just a Tool" with harm determined exclusively by how its used, or whether technical restrictions on use (Section 3.1.3) or the difficulty of using the tool influence whether it will be used for harm (Section 3.3.3). Philosopher of technology Bruno Latour and others argue against the "myth of the Neutral Tool": that the design of technological artefacts (he uses guns as a more obvious example) encode "scripts" in their design which invite certain uses and behaviors while making others harder [36, 44]. To help unearth normative conflict in discussions on software ethics, we believe it is important to discuss harms resulting from a system's implementation, the possibility for ethically questionable use, and affordances which allow the former to influence the latter.

4.3 Implications for "Ethical AI" Research: Assumptions of Downstream Control

Some companies and open source communities are wrestling with and increasingly accepting responsibility for downstream harms, as are some AI practitioners individually [61], but entrenched norms mean this is a slow and fraught process (see Section 4.1). However, mitigation strategies, for example Fairness Checklists, make reasonable assumptions about what the range of intended or possible uses are [2, 21], or weaker and often unspoken assumptions that software should not be shared, deployed or depended upon until algorithms are "sufficiently Fair". We term these **Downstream Control Assumptions**: that software producing entities *can control, know, or at least envision* how their software will be used through a mix of design intent, internal control over all the relevant features, postponing release of software, and contractual choices about appropriate customers.

For example, Google canceled its contract with the US Military to provide AI software which could be used to improve drone strike targeting (a use-based harm) after employee backlash [82] showing that Google can use contract law to exercise a fairly strict degree of control over how its proprietary software is used. This decision was politically fraught, but even before it was made, Google had a specific contractual relationship with a specific entity that it had the right to not renew, and was able to evaluate implementation harms (*i.e.*, mislabeling images) by evaluating fitness for purpose with respect to that entity's intended use.

However, as our case illustrates, assumptions of downstream control and awareness are even weaker, in both a legal and normative sense, in open source. Freedom 0 licenses *legally* dictate that contributors *may not* exercise control over how it is used, thereby enforcing the broader *norm* (see Section 4.1) that they *can and should not* be held responsible for downstream use-based harms. Open source software often has diffuse or often unknown users, and code is often freely remixed into other products [90].

Since these assumptions are so entrenched, our case suggests that "Ethical AI" research and design interventions would benefit from being explicit when making and finding ways to work effectively under loosened Assumptions of Downstream Control. "Supply chains" (the series of steps by which raw materials are converted into and delivered as a consumer product) are a construct which may help locate ethical decision making within business and community relations, and explore how different supply chain arrangements yield different outcomes. Supply chains can help reason about upstream [65] and downstream harms in [26] in offline contexts, and the UN has published actions companies should take to mitigate human rights violations in supply chains [67]. The supply chain

concept has also been transferred to software [7, 60], and software ethicists have theorized about responsibility for downstream uses of software, for example arguing that "If proper precautions are taken to limit the distribution of [hacking software], the downstream uses are constrained" [88].

This raises similar questions in other ethnographic contexts. Guides for "Responsible" use of general-purpose AI libraries often assume use(r)s can be known beforehand: guides for the general-purpose and widely-used open source ML framework TensorFlow ask "Who am I building this for? How are they going to use it?"as a crucial first step for considering ethics when using it to build other things [25]. Are TensorFlow's ethics options different or similar to the smaller use-specific project we study? In the private sector, how do far upstream actors, like ML-as-a-service companies or ML-enabling GPU manufacturers, see their responsibility and the choices available to them? Whether researchers are studying open sourced technologies or not, making explicit whether possible uses are known or unknown, and where in the supply chain possible harms or mitigations are proposed, and the limitations this may bring, can expand and strengthen AI ethics scholarship by surfacing new points of connection and action along that chain, and opportunities for ethical action under these limitations.

Funding Acknowledgement: This research was partially supported by Intel Labs.

REFERENCES

[1] 2010. The 82nd Academy Awards (2010) Nominees and Winners. http://www.oscars.org/awards/academyawards/legacy/ceremony/82nd-winners.html/. *The Academy of Motion Picture Arts and Sciences* (2010). Accessed: 2022-004-25.

[2] 2019. Google Will Not Renew Pentagon Contract That Upset Employees. https://deon.drivendata.org. *DrivenData* (2019). Accessed: 2021-07-7.

[3] 2020. A Tech Workers' Bill of Rights. https://techworkerscoalition.org/bill-of-rights/. *Tech Workers Coalition* (2020). Accessed: 2021-010-1.

[4] Christophe Abrassart, Yoshua Bengio, G Chicoisine, N De Marcellis, M Warin, Gambs Dilhac, et al. 2018. Montreal Declaration for Responsible Development of Artificial Intelligence. (2018).

[5] Christophe Abrassart and Marc-Antoine Dilhac. 2018.

[6] Henry Ajder, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen. 2019. The state of deepfakes: Landscape, threats, and impact. *Amsterdam: Deeptrace* (2019).

[7] Bilal Al Sabbagh and Stewart Kowalski. 2015. A socio-technical framework for threat modeling a software supply chain. *IEEE Security & Privacy* 13, 4 (2015), 30–39.

[8] Shaosong Ou Alexander Hars. 2002. Working for free? Motivations for participating in open-source projects. *International journal of electronic commerce* 6, 3 (2002), 25–39.

[9] Alexander Amini, Ava P Soleimany, Wilko Schwarting, Sangeeta N Bhatia, and Daniela Rus. 2019. Uncovering and mitigating algorithmic bias through learned latent structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 289–295.

[10] Mike Ananny. 2016. Toward an ethics of algorithms: Convening, observation, probability, and timeliness. *Science, Technology, & Human Values* 41, 1 (2016), 93–117.

[11] Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society* 20, 3 (2018), 973–989.

[12] Hannah Arendt. 1987. Collective responsibility. In *Amor mundi*. Springer, 43–50.

[13] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature* 563, 7729 (2018), 59–64.

[14] Justin B Biddle. 2020. On predicting recidivism: epistemic risk, tradeoffs, and values in machine learning. *Canadian Journal of Philosophy* (2020), 1–21.

[15] Nick Bostrom. 2017. Strategic implications of openness in AI development. *Global policy* 8, 2 (2017), 135–148.

[16] Marcus Burkhardt. 2019. Mapping the Democratization of AI on GitHub. A First Approach. (2019).

[17] Daniel Chandler. 1995. Technological or media determinism.

[18] Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* 63, 5 (2020), 82–89.

[19] Devin Coldewey. 2020. The RIAA is coming for the YouTube downloaders. https://techcrunch.com/2020/10/23/the-riaa-is-coming-for-the-youtube-downloaders/. *Tech Crunch* (23 Oct. 2020). Accessed: 2021-07-1.

[20] E Gabriella Coleman. 2012. *Coding freedom: The ethics and aesthetics of hacking*. Princeton University Press.

[21] Henriette Cramer, Jean Garcia-Gathright, Sravana Reddy, Aaron Springer, and Romain Takeo Bouyer. 2019. Translation, tracks & data: an algorithmic bias effort in practice. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–8.

[22] Henriette Cramer, Jenn Wortman Vaughan, Ken Holstein, Hanna Wallach, Jean Garcia-Gathright, Hal Daumé III, Miroslav Dudík, and Sravana Reddy. 2019. Challenges of incorporating algorithmic fairness into industry practice. *FAT* Tutorial* (2019).

[23] Laura Dabbish, Colleen Stuart, Jason Tsay, and Jim Herbsleb. 2012. Social coding in GitHub: transparency and collaboration in an open software repository. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. 1277–1286.

[24] Sherae L Daniel, Ting-Ting Rachel Chung, and Pratyush Nidhi Sharma. 2020. The Impact of Anonymous Peripheral Contributions on Open Source Software Development. *AIS Transactions on Human-Computer Interaction* 12, 3 (2020), 146–171.

[25] Tulsee Doshi and Andrew Zaldivar. 2020. Responsible AI with TensorFlow. https://blog.tensorflow.org/2020/06/responsible-ai-with-tensorflow.html. *TensorFlow Blog* (29 June 2020). Accessed: 2021-07-7.

[26] Jean Macchiaroli Eggen and John G Culhane. 2002. Gun Torts: Defining a cause of action for victims in suits against gun manufacturers. *NCL Rev.* 81 (2002), 115.

[27] Joel Feinberg. 1968. Collective responsibility. *The Journal of Philosophy* 65, 21 (1968), 674–688.

[28] R Stuart Geiger. 2017. Summary analysis of the 2017 github open source survey. *arXiv preprint arXiv:1706.02777* (2017).

[29] Chandell Gosse and Jacquelyn Burkell. 2020. Politics and porn: how news media characterizes problems presented by deepfakes. *Critical Studies in Media Communication* 37, 5 (2020), 497–511.

[30] Daniel Greene, Anna Lauren Hoffmann, and Luke Stark. 2019. Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. In *Proceedings of the 52nd Hawaii international conference on system sciences*.

[31] Frances S Grodzinsky, Keith Miller, and Marty J Wolf. 2003. Ethical issues in open source software. *Journal of Information, Communication and Ethics in Society* (2003).

[32] Guido Hertel, Sven Niedner, and Stefanie Herrmann. 2003. Motivation of software developers in Open Source projects: an Internet-based survey of contributors to the Linux kernel. *Research policy* 32, 7 (2003), 1159–1177.

[33] Bill Hibbard. 2008. Open source AI. *Frontiers in Artificial Intelligence and Applications* 171 (2008), 473.

[34] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–16.

[35] Daning Hu, J Leon Zhao, and Jiesi Cheng. 2012. Reputation management in an open source developer social network: An empirical study on determinants of positive evaluations. *Decision Support Systems* 53, 3 (2012), 526–533.

[36] Don Ihde. 1990. Technology and the lifeworld: From garden to earth. (1990).

[37] Sue Curry Jansen and Brian Martin. 2015. The Streisand effect and censorship backfire. (2015).

[38] Colin Jerolmack and Shamus Khan. 2014. Talk is cheap: Ethnography and the attitudinal fallacy. *Sociological methods & research* 43, 2 (2014), 178–209.

[39] Jiji. 2020. The RIAA is coming for the YouTube downloaders. https://www.japantimes.co.jp/news/2020/10/02/national/crime-legal/two-men-arrested-deepfake-pornography-videos/. *The Japan Times* (2 Oct. 2020). Accessed: 2021-07-1.

[40] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[41] Christopher M Kelty. 2008. *Two bits: The cultural significance of free software*. Duke University Press.

[42] Os Keyes, Jevan Hutson, and Meredith Durbin. 2019. A mulching proposal: Analysing and improving an algorithmic system for turning the elderly into high-nutrient slurry. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[43] Jan Kietzmann, Linda W Lee, Ian P McCarthy, and Tim C Kietzmann. 2020. Deepfakes: Trick or treat? *Business Horizons* 63, 2 (2020), 135–146.

[44] Bruno Latour. 1994. On technical mediation. (1994).

[45] Bruno Latour, Wiebe E Bijker, and John Law. 1992. Shaping technology/building society: Studies in sociotechnical change. *W. Bijker & J. Law (Eds.)* (1992), 225–258.

[46] Yvonna S Lincoln, Susan A Lynham, and Egon G Guba. 2011. Paradigmatic controversies, contradictions, and emerging confluences, revisited. *The Sage handbook of qualitative research* 4 (2011), 97–128.

[47] Zachary C Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.
[48] Kathleen M MacQueen, Eleanor McLellan, Kelly Kay, and Bobby Milstein. 1998. Codebook development for team-based qualitative analysis. *Cam Journal* 10, 2 (1998), 31–36.
[49] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
[50] Sophie Maddocks. 2020. 'A Deepfake Porn Plot Intended to Silence Me': exploring continuities between pornographic and 'political'deep fakes. *Porn Studies* 7, 4 (2020), 415–423.
[51] Noëmi Manders-Huits and Michael Zimmer. 2009. Values and pragmatic action: The challenges of introducing ethical intelligence in technical design communities. *The International Review of Information Ethics* 10 (2009), 37–44.
[52] Jennifer Marlow and Laura Dabbish. 2013. Activity traces and signals in software developer recruitment and hiring. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 145–156.
[53] Jacob Metcalf, Emanuel Moss, et al. 2019. Owning ethics: Corporate logics, silicon valley, and the institutionalization of ethics. *Social Research: An International Quarterly* 86, 2 (2019), 449–476.
[54] Matthew B Miles and A Michael Huberman. 1994. *Qualitative data analysis: An expanded sourcebook*. sage.
[55] Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1, 11 (2019), 501–507.
[56] Lewis Mumford. 1970. *The pentagon of power*. Vol. 2. Houghton Mifflin Harcourt P.
[57] Robin Murphy and David D Woods. 2009. Beyond Asimov: the three laws of responsible robotics. *IEEE intelligent systems* 24, 4 (2009), 14–20.
[58] Helen Nissenbaum. 2017. Computing and accountability. In *Computer Ethics*. Routledge, 273–280.
[59] Wonseok Oh and Sangyong Jeon. 2007. Membership herding and network stability in the open source community: The Ising perspective. *Management science* 53, 7 (2007), 1086–1101.
[60] Marc Ohm, Henrik Plate, Arnold Sykosch, and Michael Meier. 2020. Backstabber's Knife Collection: A Review of Open Source Software Supply Chain Attacks. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, 23–43.
[61] Will Orr and Jenny L Davis. 2020. Attributions of ethical responsibility by Artificial Intelligence practitioners. *Information, Communication & Society* 23, 5 (2020), 719–735.
[62] Arnold Pacey. 1983. *The culture of technology*. MIT press.
[63] Bruce Perens. 2005. The emerging economic paradigm of open source. *First Monday* (2005).
[64] Carol Righi, Janice James, Michael Beasley, Donald L Day, Jean E Fox, Jennifer Gieber, Chris Howe, and Laconya Ruby. 2013. Card sort analysis best practices. *Journal of Usability Studies* 8, 3 (2013), 69–89.
[65] Sarah Roberts. 2003. Supply chain specific? Understanding the patchy success of ethical sourcing initiatives. *Journal of business ethics* 44, 2 (2003), 159–170.
[66] Cynthia Rudin, Caroline Wang, and Beau Coker. 2018. The age of secrecy and unfairness in recidivism prediction. *arXiv preprint arXiv:1811.00731* (2018).
[67] John Ruggie. 2011. .
[68] Paul A Schewe and William O'donohue. 1993. Sexual abuse prevention with high-risk males: The roles of victim empathy and rape myths. *Violence and Victims* 8, 4 (1993), 339–351.
[69] Barry R Schneider. 1994. Nuclear proliferation and counter-proliferation: Policy issues and debates. *Mershon International Studies Review* 38, Supplement_2 (1994), 209–234.
[70] Evan Selinger. 2012. The Philosophy of the Technology of the Gun. *The Atlantic* 23 (2012).
[71] Richard Stallman. 2016. Why programs must not limit the freedom to run them. https://www.gnu.org/philosophy/free-sw.en.html. *The Free Software Foundation* (18 Nov. 2016). Accessed: 2021-07-1.
[72] Richard Stallman et al. 2021. What is free software?, The Free Software Definition, Version 1.169. https://www.gnu.org/philosophy/free-sw.en.html. *The Free Software Foundation* (2 Feb. 2021). Accessed: 2021-07-1.
[73] Luke Stark, Daniel Greene, and Anna Lauren Hoffmann. 2021. Critical Perspectives on Governance Mechanisms for AI/ML Systems. In *The Cultural Life of Machine Learning*. Springer, 257–280.
[74] Stefan Stieger and Anja S Göritz. 2006. Using instant messaging for Internet-based interviews. *CyberPsychology & Behavior* 9, 5 (2006), 552–559.
[75] Anselm Strauss and Juliet M Corbin. 1997. *Grounded theory in practice*. Sage.
[76] David R Thomas. 2003. A general inductive approach for qualitative data analysis. (2003).
[77] M Tiemann and Open Source Initiative. 2009. History of the OSI (Open Source Initiative). *Retrieved February* 4 (2009), 2009.
[78] Marcus Tomalin, Bill Byrne, Shauna Concannon, Danielle Saunders, and Stefanie Ullmann. 2021. The practical ethics of bias reduction in machine translation: Why domain adaptation is better than data debiasing. *Ethics and Information Technology* (2021), 1–15.
[79] Joachim Van den Bergh and Dirk Deschoolmeester. 2010. Ethical decision making in ICT: discussing the impact of an ethical code of conduct. *Communications of the IBIMA* (2010), 1–10.
[80] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.
[81] Travis L Wagner and Ashley Blewer. 2019. "The Word Real Is No Longer Real": Deepfakes, Gender, and the Challenges of AI-Altered Video. *Open Information Science* 3, 1 (2019), 32–46.
[82] Daisuke Wakabayashi and Scott Shane. 2018. Google Will Not Renew Pentagon Contract That Upset Employees. https://www.nytimes.com/2018/06/01/technology/google-pentagon-project-maven.html. *The New York Times* (1 June 2018). Accessed: 2021-07-1.
[83] Michael Weiss. 2010. Profiting from open source. In *Proceedings of the 15th European Conference on Pattern Languages of Programs*. 1–8.
[84] Robert S Weiss. 1995. *Learning from strangers: The art and method of qualitative interview studies*. Simon and Schuster.
[85] Meredith Whittaker. 2021. The steep cost of capture. *Interactions* 28, 6 (2021), 50–55.
[86] Rachel Winter and Anastasia Salter. 2020. DeepFakes: uncovering hardcore open source on GitHub. *Porn Studies* 7, 4 (2020), 382–397.
[87] Marty J Wolf, Kevin Bowyer, Don Gotterbarn, and Keith Miller. 2002. Open source software: intellectual challenges to the status quo. *ACM SIGCSE Bulletin* 34, 1 (2002), 317–318.
[88] Marty J Wolf, Keith W Miller, and Frances S Grodzinsky. 2019. On the Responsibility for Uses of Downstream Software. *Computer Ethics-Philosophical Enquiry (CEPE) Proceedings* 2019, 1 (2019), 3.
[89] Robert K Yin. 2017. *Case study research and applications: Design and methods*. Sage publications.
[90] Shurui Zhou, Bogdan Vasilescu, and Christian Kästner. 2019. What the fork: a study of inefficient and efficient forking practices in social coding. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 350–361.

A WHAT CAN BE DONE IN OPEN SOURCE?

Here, we centralize and summarize six tentative ideas suggested by our research for open source communities, platforms, and other researchers to explore and study the effectiveness of in other contexts in future work.

1. Take a stand on unethical uses, and enforce it. Project leaders explicitly laid out what they hope people do and do not do with their project, and why. They discouraged unethical uses by refusing to support and banning users. This strengthens norms of acceptable and unacceptable use among users in communities where leaders may have this sway.

2. Educate on project-specific harms. Project members were concerned that users who did not rely on community support might not fully think through the negative effects of unethical uses on victims. Increased victim empathy has had some effect in other contexts [68]. Explore ways to go beyond taking a stand, and proactively educate about the personal impact of unethical use.

3. Consider technical restrictions on use. Some participants believed that technical restrictions could be effective because circumvention would be technically difficult for the majority of the project's users, but leaders feared splintering the project into ones without strong norms of acceptable use. The durability and effect of such restrictions should be studied, especially as norm-setting and other measures become widespread.

4. Leverage reputational incentives. Reputational incentives partly motivated our participants to take ethical action. GitHub

contributors use features to make their own or their project's *technical* reputation more visible [23]. Platforms and researchers can investigate affordances to make downstream uses and impact of code more visible, which more closely tie the reputation of a project and those who work on it to the impact that project has.

5. Platforms should publish and consistently enforce policies that consider downstream uses. Platforms supporting open source development and dissemination do not always share the absolutist view that the right to use software for any reason should be protected. Google, GitHub and Discord all took action against the project we study or its users. However, it is not clear that the project violated GitHub's content restrictions,[4] and similar projects which actively promoted pornographic Deepfakes reportedly did not face similar action, suggesting the need for clear and consistently enforced standards.

6. Publicize and study Ethical Source licenses. In a more recent innovation, some have designed alternative licenses to control the ethical impact of downstream uses, known as the "Ethical Source" movement. Researchers should study their adoption and wider effects, and platforms such as GitHub can educate on the availability of these licenses, as they do for more traditional Free and Open Source Software Licenses. This will challenge the thinking that leaves Freedom 0 as default, and clarify misconceptions over the implications of license choice.

[4] See: https://docs.github.com/en/github/site-policy/github-acceptable-use-policies