bark 🐕

October 22, 2021

Senators Warren, Markey, and Blumenthal
Washington, D.C. 20510

Dear Senators,

I want to thank you for your recent focus on children's online safety, including investigating how social media platforms' decisions can have negative impacts on children. I founded Bark over six years ago with the mission of making the internet a safer, better place for kids and their caregivers, and I have been continually dismayed by platforms' repeated choices to prioritize profit at the expense of children's safety.

Now that there is more widespread attention to the harms that children face from social media platforms and a clear conclusion that social media platforms cannot be trusted to provide safe environments for children, the question becomes what should be done. There are fundamentally only a few options: disallow children from using social media altogether, demand the platforms do better, and/or enable families with the power of choice and the tools they need to help children have safe experiences online. Disallowing children from using social platforms is impractical, and could have some downsides as it is possible there are positive benefits for children using social media. Given the deeply ingrained misalignment between social media platforms' incentives and children's safety, expecting the platforms to provide sufficient solutions (even with regulation) is an incomplete-at-best solution. We believe that empowering caregivers to help guide their children as they navigate technology is a critical component of the solution. Several platforms that are used ubiquitously by children do not provide basic data portability that would allow caregivers to use safety products for their families. This is analogous to an auto maker who intentionally makes cars that do not have seatbelts or air bags, AND who also takes active steps to prevent their customers from installing those basic safety tools themselves. Consumers should be able to exercise choice in using standard safety tools to maintain safe experiences on the platforms they use.

I also thank you for asking the right questions to ensure that those in the school safety space (as well as in other domains, such as the financial, medical, and communication sectors) are doing everything possible to provide effective and equitable solutions. At Bark we have been thinking deeply about these topics for the last six years, refining our methods and approaches by engaging and working with a wide array of advisors, non-profits ████████████████ ██ government agencies ████████████████ tech companies ████████████████ ████████, and academic experts to ensure we're considering all perspectives in our approaches, and have invested heavily in providing products that advance children's online safety while prioritizing equity and privacy.

bark 🐕

As an industry we need to continue to work diligently to provide the best products and services possible, and to consider simultaneously many related concerns, such as a school's obligation to provide safe technology to students, the rising rates of negative-but-preventable outcomes for teens/tweens, and the privacy of students. I believe the goal should be to have transparent conversations that drive the best possible outcomes for children.

Our primary product is Bark for Families, which is a suite of tools to help families keep children safe on personal devices and social media accounts. Following the horrific attack at Parkland High School, we created and launched our Bark for Schools product as a give-back to help schools reduce preventable tragedies. We make this product free to any US K-12 school that wishes to use it.

Bark for Schools provides free DNS and Chrome-based web filters, which can help schools with CIPA compliance. Schools can also choose to connect Google Workspace and Microsoft 365 accounts to Bark for Schools, and content for those connected accounts is analyzed algorithmically by a suite of deep learning-based models, which look to identify situations of potential bullying, violence, self-harm, etc. Unlike less sophisticated solutions, Bark does not simply flag keywords, which is a naïve approach that results in unneeded alerts, but rather uses state-of-the-art deep learning approaches to determine more holistically if there is an issue. When an issue is detected, alerts are sent to the appropriate administrator(s) at the school (as designated by the school). If the school has chosen to use our free Parent Portal feature, and the relevant parent has opted in to its use, the relevant parent will also be sent an alert. When an issue is flagged by our system as an imminent, severe risk (ranging from school shooting threats to imminent suicidal situations), it is escalated for human review by our 24/7 team of trained reviewers who can call school administrators and/or escalate the situation to law enforcement and/or NCMEC as needed. In all cases, only the part of a conversation that has been flagged is shown in the alert – we intentionally do not provide any way for a school or parent to read messages that are not flagged as potentially problematic. Additionally, only the message content is analyzed; the algorithmic models do not have access to any structured personal student data (such as name, age, gender, race, location, etc.).

Our free Bark for Schools product is currently used by over 2,900 schools/districts, which corresponds to over 5.5 million students. Less than 100 of these schools/districts have purchased our relatively new Bark for Schools+ ("Plus") product, which adds a handful of additional features for a retail price of $2/student/year. There is no difference in our data privacy policies between the free and Plus products.

Given our experience alerting caregivers and school administrators to severe, imminent issues, we strongly urge that you consider creating and funding a standardized, 24/7 system of mental health emergency workers. At Bark, we currently send an average of 85 alerts every day for imminent self-harm/suicidal situations, and if we are unable to reach a school administrator

or the child's caregiver directly there are limited options to reach someone who can provide help for a child in the moment of crisis. In most places the only option available in this situation is to call local law enforcement. And while we have witnessed many lives be saved by police in these situations, unfortunately many officers have not received training in how to handle such crises, and irrespective of training there is always a risk that a visit from law enforcement can create other negative outcomes for a student or their family.

At a high level, we all recognize there is no such thing as a perfect system – every step forward we take as society has some negative impact on some amount of people. Every medical treatment that benefits many people causes some unintended harm to some percentage of the population. Every law passed to help the masses will unintentionally hurt others. That same real-world trolley problem exists with any system, including those using AI. Because of this, the primary question we should seek to answer is "Do the benefits of a given system outweigh the tradeoffs?" followed by "Are the actors in that system actively trying to continue to shift the balance of benefits vs. tradeoffs in an increasingly positive direction?"

To that first question, across 5.5 million students, on accounts/devices owned/issued to those students by schools, Bark for Schools has alerted school administrators to:

- 15,161 instances of imminent suicide/self-harm
- 31,214 instances of severe violence
- 290,694 instances of nudity
- 83 instances of student conversations with child predators
- 28 alerts of sextortion
- 102,763 instances of severe hate speech
- 454,300 instances of students buying/selling illegal drugs
- 384,780 instances of severe depression
- 756 instances of radicalization by hate groups
- 134,272 instances of severe bullying
- 5,913 instances of negative body image

(Note: While we keep counts of alerts by type, because we do not retain all of the underlying data we are not able to report on activities analyzed by Bark for Schools by time of day or by day of week.)

All of the above alerts represent opportunities for a trusted adult to help a child in a moment of crisis. It is of course possible that for some of these situations there may have been awareness of the situation from other sources (such as a report from a student or teacher), but the feedback we have received from schools and parents is that most issues would go undetected without the alert from Bark. We strongly endorse continued educational efforts to encourage students to involve a trusted adult in severe situations, as a multi-faceted approach will

certainly yield the most positive outcomes in reducing the widespread issues that children face today.

As highlighted in the Center for Democracy & Technology report which you referenced in your September 29, 2021, letter, "Teachers, parents, and students largely report that the benefits of monitoring software outweigh the risks." While the massive benefits are clear, we must be diligent to ensure that the processes by which we generate these potentially life-saving alerts continue to improve, as we consider the full spectrum of interests impacting the welfare of families and students.

As your letter also points out, there is a terrible history of bias in school discipline issues irrespective of usage of any AI-based solution, as new research confirms. While any system, including AI-based solutions, inherently have some bias, if implemented correctly AI-based solutions can substantially reduce the bias that students face. Whereas teachers and school administrators make decisions every day that are affected by their individual biases, a correctly implemented AI approach spreads the decision-making across a wide set of training data and – more importantly – train*ers*. Academic research has long shown that using machine learning models trained on human-annotated data counteracts individual preferences and biases that can lead to inconsistent and incorrect human decisions. By learning from many humans' judgments, models are able to ignore individual biases and noise to learn the "wisdom of the crowd." At Bark, we go further to explicitly require exact agreement in annotation across multiple independent reviewers to be included in our training datasets. Coupled with a focus on both diversity in hiring and bias reduction training, we can provide a product with substantially less bias than school personnel.

Another advantage of AI systems is that, where bias exists, it can be more easily detected than when those biases exist in individual decision makers (e.g., school personnel) or social institutions. Moreover, where biases are detected, AI training algorithms also provide relatively quick and transparent means to address them. For example, a recent study found that computing simple training weights can eliminate racial biases found in popular hate-speech training datasets. Relatedly, in public health policy, recent research from Harvard demonstrates how constraints can be applied to the objective function in AI models to systematically enforce fairness in distributing HIV protection interventions. These findings showing rapid progress in addressing algorithmic bias led one of the leading researchers on algorithmic bias, Sendhil Mullainathan, to conclude that while algorithmic biases surely exist, they are drastically easier to fix than comparable biases in people.

The Center for Democracy and Technology reports that algorithmic systems stand to significantly benefit students' learning environments, well-being, and opportunities, and lays out a clear list of actions that should be taken to mitigate any negative tradeoffs: "To address these considerations, education leaders and the companies that work with them should take

the following actions when designing or procuring an algorithmic system." Bark's products and approaches fully align with the guidance from that list. We believe in transparency and have spoken publicly about the potential tradeoffs of using AI, and look forward to evolving research in the space.

We agree with asking questions that seek to ensure proper use of any system or technology. We also believe that it is important also to consider the alternatives to the usage of any system. If the alternative to Bark for Schools is to rely solely on awareness of issues coming from students, teachers, and principals, history suggests that those decisions will, in comparison to platforms like Bark for Schools, be more prone to individual biases and noise, subject to systemic biases that are harder to detect and measure, and more challenging to de-bias.

Beyond the question of bias, it is also necessary to weigh the tradeoff of a high amount of negative-but-preventable student outcomes by not using a system such as the free Bark for Schools product. Given the rising rates of suicide attempts, where more than 1 in 10 high school girls attempt suicide each year, there is a massive cost in turning a blind eye to warning signs. The American Academy of Pediatrics, the American Academy of Child and Adolescent Psychiatry, and the Children's Hospital Association have jointly declared a National State of Emergency in Children's Mental Health. We very much agree with your concern that more collective effort be focused towards groups who bear disproportionate amounts of negative outcomes. For instance, the LGBTQ+ community experiences a significantly higher frequency of both suicide attempts and bullying, and without awareness of these situations, caregivers and school administration are unable to intervene. In the last three years, Bark for Schools has sent over 15,000 alerts for situations of imminent self-harm, and over 134,000 alerts of extreme bullying, providing school administration and caregivers opportunities for intervention that would not otherwise have existed.

While we have invested heavily in mitigating potential tradeoffs of our products, we will continue to work aggressively to get even better as we learn and as technology evolves. We also recognize that additional research may provide more guidance. For example, even absent usage of any technology, schools – comprised of humans who all have inherent biases – face challenges with making biased decisions in issues such as student discipline, and it would therefore be important to conduct a controlled study that compares the overall amount of bias in educational settings that utilize correctly implemented monitoring software versus those that do not. If done properly, usage of AI should yield overall lower amounts of bias in an educational setting, because the effect of an individual teacher's or administrator's bias is substantial relative to a system intentionally designed to minimize bias in the system.

We utilize several tools and procedures to test for and minimize bias. We regularly test our models for bias against references to protected classes and behaviors that could unnecessarily (i.e., without any abuse or harmful behavior present) reveal personal details of a student to

parents or administrators. With respect to race and gender, our models do not flag otherwise inert, first- or second-party references. With respect to sexual preferences and disabilities, our models dissociate inert, first-party references (e.g., "I am gay," which would not be flagged) from second-party statements which are more common in students' bullying (e.g., "You are so gay," which would be flagged). (Note: Because we do not collect any student sexual identity or preference information, we cannot analyze results by protected class.)

Our audit practices were borne from a regular review of our data science processes and models using an industry-standard Data Science Ethics Checklist, which is a framework for analyzing the ethical considerations of data science projects, from data collection/storage through analysis, modeling, and deployment.

Many studies on hate speech and abusive text detection rely on datasets that have been shown to be biased. Research has shown that, even with these biased datasets, bias can be mitigated through careful training processes, for example, such as reweighting the data to address racial biases. At Bark, to avoid the bias in freely available abusive-text datasets that are used in most academic research (e.g., Davidson et al., 2017), we rely on first-party data annotations from a team of highly trained annotators. This ensures that our abuse guidelines are consistently applied at a higher standard than crowdsourced online datasets (e.g., Founta et al., 2018). It also ensures that our training data are sampled from the population of data on which our models are applied after deployment, avoiding generalizability concerns that can hinder model performance and increase false positive alerts.

Model training data are stripped of all personally identifiable information ("PII"), using technology we have developed using a combination of string-pattern matching and probabilistic classifiers, and are then annotated by the Bark annotation team, all of whom are employees of Bark, located in the United States. Each day, we automatically rate the performance and reliability of the annotation team across each of the abuse categories and languages scored by our models. Annotators who show declining reliability are given additional training to improve their performance. For a data point to be added to the training dataset, multiple data annotators must be in agreement on the annotation, which mitigates individual bias from being learned by the model. Disagreements on annotation are manually investigated to explore any possible improvements in process or training.

Training data is then augmented to increase model robustness and reweight samples to address any over- or under-sensitivities observed. Once ready, data are iteratively used to train the models that perform real-time abuse detection. Model accuracy is evaluated first on a holdout dataset for overall performance and for performance in predetermined data slices (e.g., in high-severity violence- and suicide-related texts) and then benchmarked against other model release candidates. Prior to model deployment, all candidate models are evaluated using an additional set of pass/fail tests, including texts that must *not* be flagged by the model (e.g., first-

party statements of sexual preference) and statements that *must* be flagged by the model (e.g., school shooting threats, statements that indicate imminent self-harm). Candidate models that do not pass these tests are not released and are introspected through further model research.

Bark for Schools does not collect any structured student data such as grades, attendance data, or PII data listed at https://studentprivacy.ed.gov/ferpa other than name and email. We do not share any student information with any third party, other than with law enforcement in severe cases. We do store aggregated counts of types of things flagged, so that we may produce stats/trends that are useful publicly (e.g., change in rates of cyberbullying over time).

Analyzed content that is not flagged for any issue (which accounts for about 99.7% of content analyzed) is automatically purged within 30 days; content that is flagged for an issue is stored until the school administration completes its review of the issue.

In terms of data security, we are SOC 2 Type 2 compliant with third-party audits, and undergo regular penetration tests. We secure all of our data within an encrypted database, including backups. Our server infrastructure is located in highly secured physical data centers that use a centralized bastion host, which is monitored to detect unwarranted access or activity. All web browser sessions use and require SSL encryption.

We wholeheartedly agree that schools should clearly make students and their caregivers aware of the use of any security/threat detection tool, whether security cameras or software-based products. While we have no ability to dictate how a school communicates with its students' parents, we encourage schools to be fully transparent with students and their parents about the usage of our product. For a school to use our Bark for Schools product, a school must "represent and warrant that [they] have the legal authority to access, monitor, review and store online interactions and other communications to and from such Covered Account, including without limitation, all legally required consents from a parent or legal guardian of a Covered Account for Bark to provide the Services in accordance with these Terms and our Privacy Policy."

Schools may turn off the use of Bark for Schools for any set of students that they choose. Given schools each have their own technology usage policy agreements between them and students/guardians, the process of opting out of usage of Bark for Schools is necessarily between a student/guardian and the school. We are unaware of any students/guardians requesting to opt out, but we would likely not have visibility into those requests should they exist.

It is our belief that schools should never have any visibility into students' communications made on students' personal devices/accounts, and as such Bark for Schools only monitors devices and accounts owned/issued by the school, the usage of which is governed by an agreement between the school and the student. We support the Supreme Court's recent ruling

that protects students' right to off-campus speech in many contexts; however, to address the question in your letter, this ruling seems unrelated to a school's usage of Bark for Schools. In that ruling the Court made clear that schools remain able to enact discipline for off-campus speech when such speech includes serious or severe bullying or harassment targeting particular individuals, threats aimed at teachers or other students, the failure to follow rules concerning lessons, the use of computers, participation in other online school activities, breaches of school security devices, etc. Also, this ruling does not seem to have any bearing on the method/medium by which a student chooses to voice that speech, nor guarantee privacy for speech made through a device/account owned by the school, nor does Bark for Schools make any discipline-related decisions or recommendations. It is always the decision of school administration as to whether to enact any disciplinary measures, irrespective of whether their awareness of a situation comes from Bark for Schools, from a teacher, or from another student. Off-campus speech made on a school owned/issued device/account seems analogous to a student walking on to school premises in the middle of the night – they are likely not allowed to do so (just as technology usage agreements often do not allow students to use the devices for personal use) and even if they are allowed to do so, this does not mean that the student should expect privacy from the school's security camera.

Thank you again for shining a light on the massive challenges children face today, for your efforts to ensure technology is being used appropriately, and for weighing the benefits of such technology as compared with any unintended tradeoffs. We look forward to continuing this dialogue with all stakeholders concerned with children's online safety and the ongoing efforts to ensure technology is maximally beneficial and equitable for all children in our country.

Best regards,

Brian Bason