CASC | Coalition for Academic Scientific Computation

CASC Response to Request for Comment NTIA-2024-0004

V2 1/14/2025

1. What are the potential benefits of developing national-level ethical guidelines for researchers collecting, analyzing, and sharing pervasive data?

Research plays a critical role in the age of big data, AI, and social media, as these technologies increasingly influence society. Developing national ethical guidelines for researchers can provide several benefits:

- AI Workforce Training: Establish a foundation for training professionals to ethically develop and manage AI systems.
- Public Education: Offer balanced insights beyond sensationalized AI success stories or fears of an AI takeover.
- Validation and Compliance: Enable the creation of testing and validation frameworks to ensure tools and environments comply with ethical standards.
- Legislative Support: Inform the development of laws to safeguard public interests.

Given the nature of pervasive data, these guidelines must extend beyond academic and government research. They should address data collection, governance, and the operational processes of service providers to ensure ethical practices across all stakeholders.

2. What are the potential drawbacks of developing national-level ethical guidelines for researchers collecting, analyzing, and sharing pervasive data?

A potential drawback is that service providers and businesses involved in developing and deploying pervasive data and AI tools may perceive ethical guidelines as a precursor to legislation, potentially viewing this as a threat to their short-term profitability.

To address this concern, it’s essential to frame the guidelines within a context that highlights their long-term societal and economic benefits. Emphasizing how ethical practices can foster trust, innovation, and sustainable growth may help gain broader acceptance and adherence.

4. What are some existing barriers to accessing pervasive data?

Pervasive data is primarily collected, owned, and managed by service providers of social platforms. Open access to this data is highly challenging because its content and structure are deeply intertwined with the intellectual property (IP) of the service providers. Unlike medical research, where IRB-approved studies grant limited data access (e.g., specific data points for a set number of patients from electronic medical records), this approach is no longer sufficient for modern research needs. To achieve meaningful insights and correlations, researchers and their AI tools require comprehensive access to extensive datasets. The traditional model of limited data access does not meet the demands of contemporary research in medicine or pervasive data.

While researchers within service provider companies may have appropriate access to such data, their findings could create tension with management if the results cast the company in a negative light. For independent researchers from academia or government, access to pervasive data is even more restricted, as companies are reluctant to share their data due to IP concerns. This issue has already been documented in studies examining social media behaviors.

6. Consent and autonomy are key principles in human subjects research ethics. However, users of online services may be required to divulge certain personal information and/or have no ability to freely make decisions about its use. How should researchers working with pervasive data consider consent and autonomy?

Users of online services are often required to provide personal information, with limited ability to control how it is used. The vast scope and complexity of pervasive data make the traditional concept of consent difficult to apply.

Two primary categories of pervasive data collection impact research:

1. **Closed, authenticated environments**
   These include services such as online banking, shopping platforms (e.g., Amazon, Dillards), dating apps, and closed communication platforms (e.g., Teams, Slack). In these environments, users authenticate their identities and accept terms of service, expecting privacy in their interactions with service providers or defined groups. Users in this category are customers who either pay directly for the service or indirectly through their transactions.
2. **Massively open environments**
   Platforms such as Facebook, YouTube, TikTok, and Twitter/X operate in more open settings, where shared content is often publicly accessible by design, even though visibility controls may exist. These platforms thrive on advertising revenue, and users act as consumers rather than customers. Given the openness and large-scale propagation of data, users cannot reasonably expect privacy.

The distinctions between these two environments necessitate different approaches to consent and autonomy in ethical research guidelines:

Case 1: Closed, authenticated environments

Ethics guidelines for these environments could follow one of two approaches:

- **Obtaining consent:** Researchers could seek explicit consent from users before data collection. However, this approach introduces bias, as not all users provide consent, and those who do may not represent a statistically valid sample of the population.
- **Right to be forgotten:** Similar to GDPR principles, users could be given the option to request the removal of their data.

The context here aligns somewhat with the privacy standards of telephone networks, where conversations are considered private and law enforcement must obtain a warrant to access them. Encryption can provide an added layer of protection if the network's trustworthiness is in question.

Case 2: Massively open environments

Given the societal impact of these platforms, research in this category is critical. Ethical guidelines in this context should:

- **Permit research without consent:** Given the public nature of the data and its widespread dissemination, obtaining consent is impractical. Similarly, implementing a "right to be forgotten" is often unfeasible.
- **Focus on service providers' practices:** Instead of solely targeting researchers, guidelines should emphasize the responsibilities of service providers, including transparency about what data is collected and how algorithms influence user behavior.

Additional Recommendations

Ethical guidelines should also address organizational responsibilities. Organizations that require or encourage employees to use services collecting pervasive data should limit such use to paid, closed environments that ensure privacy. They should explicitly prohibit work-related activities on massive-scale social media platforms to mitigate ethical and privacy risks.

By tailoring ethical approaches to these two distinct categories, researchers and policymakers can better address the complexities of consent and autonomy in pervasive data research.

7. What ethical issues and risks to privacy and other rights, and mitigation strategies, should be considered during the research design phase?

The primary focus of research on pervasive data lies in understanding the provenance and origin of the data. To derive meaningful insights, research often requires access to comprehensive datasets, enabling the identification of hidden correlations and connections through advanced tools, such as AI. This inherently ties the research design to the operational processes of service providers collecting the data—an area largely beyond researchers' direct control.

Key considerations for research design:

Classification of data environments:
Ethical guidelines must distinguish between the two primary types of pervasive data (as outlined in question 6):

Type 1: Closed environments, where users/customers have an expectation of privacy. Examples include banking services, online shopping platforms, and private communication tools. Research involving these datasets must account for privacy expectations and align with stringent ethical standards.
Type 2: Open environments, such as social media platforms, where users/consumers cannot reasonably expect privacy. Research here must address ethical concerns related to the public nature of the data and its potential for misuse.
Ethical guidelines tailored to data type:

For Type 1 data, researchers must ensure privacy safeguards and consider informed consent or similar mechanisms, such as a "right to be forgotten," to uphold user autonomy.
For Type 2 data, the focus should shift to the service providers' operational processes, emphasizing transparency about data collection and algorithmic practices. Researchers must carefully evaluate the societal implications of their work to mitigate potential harm.
Use of participatory research:

While participatory research (involving active engagement with users) can be incorporated into research plans for both types of data, it may not fully leverage the potential of pervasive datasets. To achieve statistically meaningful outcomes, researchers often need to deploy AI agents that interact with users and collect large, diverse datasets. This introduces additional ethical considerations, such as ensuring fairness, avoiding bias, and preventing exploitation.

8. What are the risks and mitigation measures related to pervasive data acquisition and access?

b. Inherent Bias in Existing Data
Pervasive datasets collected over decades by major Internet tech giants already contain inherent biases. This is because not all populations have had equal access to the necessary technology, such as computers and smartphones, to actively participate in the digital world. These disparities have resulted in datasets that are not fully representative, leading to biased

outcomes. To mitigate this, research using limited subsets of pervasive data must be carefully designed to address specific questions relevant to the selected subset. Researchers should not assume that findings from such limited datasets can represent the broader population or the entirety of pervasive data.

c. High Risk of Re-identification
The risk of re-identification is particularly significant when working with pervasive data due to the sheer volume of information and the integration of disparate data sources. As AI technology evolves, it will increasingly enable reliable re-identification, even in datasets that have been de-identified. To address this, ethical guidelines should explicitly warn researchers about the limitations of de-identification methods. Researchers must take additional precautions, such as minimizing the collection of personal identifiers, adopting robust anonymization techniques, and regularly assessing re-identification risks.

d. Controlled Access to Data
Providing controlled access is likely the most effective strategy for ensuring researchers can work with pervasive data responsibly. However, service providers are unlikely to grant meaningful access without formal agreements. These typically include non-disclosure agreements (NDAs) and data use agreements (DUAs), which outline the terms of access, data usage, and confidentiality requirements. To mitigate risks, these agreements should emphasize transparency, limit the scope of permissible uses, and require compliance with ethical guidelines and privacy safeguards.

By addressing these risks and implementing appropriate mitigation measures, researchers can responsibly access and utilize pervasive data while minimizing harm and bias.

10. What are the risks to privacy and other rights related to the dissemination and archiving of research outputs? What mitigation measures exist?

b. Providing controlled access to pervasive data is likely the most practical approach for enabling researchers to work with these datasets. However, service providers may require researchers to sign data use agreements (DUAs) that grant the providers the right to review research findings before publication. This review process is typically intended to ensure that no proprietary information or intellectual property is inadvertently disclosed. However, it introduces a significant conflict of interest, particularly if the research results portray the service provider in a negative light, potentially impacting their reputation or financial performance.

To address this issue, ethical guidelines could offer a framework for achieving a balanced compromise among the service provider, the researcher, and the broader societal benefit. These guidelines should aim to ensure that research findings are published in an unbiased manner while respecting intellectual property rights and the public's right to knowledge. Transparent agreements that clearly define the boundaries of permissible reviews and safeguard against undue influence can help mitigate this challenge.

c. Reproducibility of research results faces additional challenges in the context of controlled access to pervasive data. Due to the requirement for researchers to sign non-disclosure agreements (NDAs) and DUAs to access data, others attempting to replicate the research will need to navigate the same legal and procedural barriers. This can restrict the ability of independent researchers to verify findings, thereby undermining the principle of reproducibility. Ethical guidelines should emphasize the importance of data-sharing frameworks that promote transparency and reproducibility while protecting intellectual property and user privacy. Solutions such as anonymized datasets, synthetic data, or shared access environments could be explored to address this issue.