UNIVERSITY OF TORONTO | SCHWARTZ REISMAN INSTITUTE FOR TECHNOLOGY AND SOCIETY

October 16, 2020

Manager of Access and Privacy Strategy and Policy Unit
Ministry of Government and Consumer Services
Enterprise Recordkeeping, Access and Privacy Branch
134 Ian Macdonald Blvd.
Toronto, Ontario M7A 2C5
Access.privacy@ontario.ca

Via email

Dear Ministry of Government and Consumer Services,

Re: Public Consultation - Reforming Privacy in Ontario's Private Sector

The Schwartz Reisman Institute for Technology and Society (SRI) is a new landmark institution at the University of Toronto dedicated to creating and integrating world-class research across sectors and disciplines to deepen our understanding of how powerful technologies like AI can serve the common good. SRI fosters fundamental research with cross-disciplinary teams and focuses on bringing these insights into new solutions for society. Privacy and data governance models are a primary area of our research. We are interested in developing the next generation of methods and models of regulation that can rigorously protect important rights and interests as well as unlock important data uses that are in the public interest. We bring a multidisciplinary lens to these problems, convinced that novel collaborations between researchers that include machine learning scientists, cybersecurity researchers, legal scholars, political scientists, ethicists, and economists will enlarge our solution space. Additional information about SRI is available here.

We would like to offer some comments in response to your consultation document, "Ontario Private Sector Privacy Reform: Improving private sector privacy for Ontarians in a digital age". In addition, we would like to extend an invitation to follow-up with us about any of our research that might be helpful to you in this project. We have an active research community that includes our Director, Leadership Team, Faculty Affiliates, and Fellows and also a "Solutions Stream" that uses design-thinking and innovative methods to create novel generative spaces for developing concrete ideas for practical implementation. An example of our work in this domain can be seen here.

Our comments below stem from a more general critique that we have developed of existing legal models for the protection of privacy. Legislative models based upon the Fair Information Practice Principles (FIPPs) -- such as PIPEDA or the GDPR -- are centered upon the data flow between an individual and an organization. In the 21st century these models have increasing difficulty in responding to what has become a complex data ecosystem, with many third parties who have no direct relationship with the data subject, where mass digitization means that "data" is no longer limited to text fields collected in formal processes, and where advanced data analytics allow for methods of data processing that were not contemplated by these legislative models. These new complex data flows are often characterized by a lack of transparency that undermines trust and accountability in the data ecosystem. As we outline in more detail below, these features of opacity and complexity call into question some of the features of existing legislative models and suggest the need to think about new methods of regulation. As Ontario contemplates the degree to which it should follow existing models, or depart from them, we would like to offer our critical perspective.

Opaque Data Flows

The lack of transparency in data flows is most often discussed in terms of how this affects informed consent and often focuses on the defects of privacy policies. Increasing transparency is an important goal but it is not clear that the way to meet this goal is through our standard regulatory toolkit, including improvements to the clarity of privacy policies. Such an approach remains too rooted in the idea that a privacy policy is a document meant to enhance the meaningful consent of an individual consumer with a focus on whether a consumer can understand the policy. Instead, Professors Lie and Austin argue that "we need to stop thinking about privacy policies in terms of whether consumers read and understand them and instead treat them as self-reporting mechanisms for data practices and then regulate them in the way we do other important disclosures such as financial disclosures or tax reporting -- by imposing standards and auditing for compliance."[1] Their work suggests that we can use AI to automate

the reading and classifying of privacy policies in order to support the work of *regulators* by helping them to both 1) understand information practices at scales otherwise difficult to obtain, and 2) automatically map privacy policies to data flows. In one project, they created a research tool that could detect inconsistencies between a mobile application's privacy policy and its potential data flows. When tested on 700 applications, they found

[1] David Lie, Lisa M. Austin, Peter Yi Ping Sun, and Wenjun Qiu, "Automating Accountability? Privacy Policies, Data Transparency, and the Third Party Problem" 2020, p.40 (On file with authors.)

inconsistencies in 60%, and that 80% of this problem was due to third party code (advertising, analytics).[2]

The consultation document discusses the idea that where there is “clear” transparency about an organization’s practices, consent might not be necessary. There are two problems with this. First, from a consumer perspective so many of the problems with consent regimes are at root problems with transparency -- consumers do not understand how their data is being collected, used, or disclosed. Continuing to try to simplify privacy policies and expect consumers to read them is unlikely to work.[3] If privacy policies were drafted in a manner that optimized our ability to create accurate AI models then we could automate the reading and classifying of privacy policies and build consumer- friendly tools on top of this, as well as better tools for auditing compliance.

The second problem with requiring transparency but not consent is that this is not as protective of consumer interests as some other models that Ontario could consider. Without some further constraint on the purposes for data processing (in addition to other constraints such as data minimization), transparency alone risks further entrenching the “take-it-or-leave-it” situation that is eroding consumer trust in the data ecosystem. An important alternative model is one that rests not on consent but that would require that the collection, use and disclosure be *reasonable*. Professor Austin’s work has shown that many decisions of “implied consent” under PIPEDA were, in fact, based on a reasonableness standard -- consent was implied in contexts where the data flows were found to be “reasonable” and there was proper notification.[4] This is similar to how BC and Alberta permit personal information about employees to be collected, used or disclosed *without* consent so long as there is notification and the collection, use or disclosure is “reasonable” in the context of the employment relationship.[5] Both of these approaches are similar to the approach taken in the GDPR, under which data may

[2] Lisa M. Austin, David Lie, Peter Yi Ping Sun, Robin Spillette, Mariana D’Angelo, and Michelle Wong. *Towards Dynamic Transparency: The AppTrans (Transparency for Android Applications) Project*. 2018. Online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3203601.

[3] Gillian K. Hadfield, Robert Howse and Michael J. Trebilcock, "Information-Based Principles for Rethinking Consumer Protection Policy" Journal of Consumer Policy Vol 21 pp. 131-169 (1998).

[4] Lisa M. Austin, “Is Consent the Foundation of Fair Information Practices? Canada’s Experience Under PIPEDA” (2006) 56 UTLJ 181

[5] *Ibid.*

be lawfully processed without consent on a number of grounds, including where doing so furthers legitimate interests.[6]

Complex Data Flows

Complexity in data analysis has led to pressure on the idea of "personal information" as the central organizing idea in a regulatory model. As the consultation document indicates in its discussion of "deidentified" and "derived" data, this idea of "personal information" no longer easily maps onto our data processing practices.

We think that regulatory models need to move away from approaches that emphasize categories of data. This is already a problem in relation to "personal information" in Canadian data protection law statutes -- the statutes adopt a binary approach (yes-then-regulated/no-then-not-regulated) when the research community consistently shows that whether information is identifiable is better conceived of in terms of a spectrum of risk.[7] Our regulatory models should regulate information about people, not "personal information", so that it is clear that information that you describe in the consultation document as "deidentified" or "derived" is also regulated. We should then impose *differentiated* obligations that more closely align with the nature of the data processing and levels of risks at issue. This would also provide a better general framework for addressing some of the issues that arise when data becomes part of our public infrastructure, as in the "smart city" context. In such cases, focusing on whether the data is "personal information" or not and then regulating it according to the FIPPs model does not adequately address the many important interests at stake.[8]

One of the difficulties of focusing on categories of data, including "deidentified" data, is that it encourages focusing on features of the data alone rather than the analysis that is done on the data or the computing environment within which this is done. The focus should be on encouraging a variety of methods to manage re-identification risks rather than on deidentification. Lawyers sometimes refer to all of these types of methods as a kind of "contextual" deidentification. But such an approach is trying to shoehorn new approaches into old categories and only creates confusion and impedes the development

[6] GDPR, Article 6.

[7] Lisa M. Austin, "Reviewing PIPEDA: Control, Privacy and the Limits of Fair Information Practices" (2006) 44 CBLJ 21.

[8] Lisa M. Austin, "Data Trusts and the Governance of Smart Environments: Lessons From the Failure of Sidewalk Labs' Urban Data Trust" (2020) (on file with author).

of clear technical standards. Furthermore, a machine learning model may leak some types of information while being blind to other types depending on the objective the model was trained for. A focus on the data (is it deidentified?) and not the *model* leaves out important questions about both sources of risk methods to mitigate those risks.[9]

Increasingly, privacy-preserving data analysis in computer science is concerned with making the *analysis* itself private -- rather than manipulating the data -- in order to be able to operate on the raw data directly. A popular framework for doing so is called differential privacy. When learning with differential privacy, one can provide statistical guarantees demonstrating that the contribution of an individual to the machine learning model is well understood and limited to reinforcing patterns already found in the data of other individuals. In contrast, approaches to privacy that focus on manipulating the data to ensure that the data is "deidentified" can harm the ability to use machine learning but do not provide privacy guarantees that are as strong.

Privacy can also be protected through methods of data analysis that provide a secure means of allowing one party to perform computations on data held by another party without revealing to them the underlying raw data.[10] This is also very different from sharing "deidentified" data with another organization as it does not involve sharing and the privacy of the raw data is managed through a variety of techniques that manage the risk of re-identification within that compute environment.

These new methods of analysis and means of creating trusted computing environments call into questions other traditional regulatory categories from Canadian data protection legislation. Most of this legislation uses the terms "collect," "use" and "disclose" to refer to data flows. However, consider the following scenario. Suppose organization A wants to share personal information with organization B so that B can link this with its own information and perform some computation. There are methods of analysis on data where an output can be computed while the underlying raw data remains hidden, such as Secure Multiparty Computation and Homomorphic Encryption. If A and B use these techniques, then is A still disclosing personal information to B? Is B using personal information? It seems like neither is true; but having this unregulated is also problematic. The categories simply do not fit the techniques.

[9] Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar, "Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data" (https://arxiv.org/abs/1610.05755); Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, Úlfar Erlingsson, "Scalable Private Learning with PATE" (https://arxiv.org/abs/1802.08908).

[10] Lisa M. Austin and David Lie, "Safe Sharing Sites" (2020) 94 NYU L Rev 581.

Another feature of complex data flows is that it makes data erasure difficult in some contexts. For example, if Ontario decides to recognize a right to data erasure, it should clarify the extent to which this right includes requesting that one's personal information be removed from a machine learning model that might have been trained with it.[11]

New Approaches to Regulation

Our existing approaches to privacy regulation in Canada are rooted in the regulatory techniques of the twentieth century. With the explosion of mass digitization and globalization, these approaches are struggling to keep up with the complexity and speed of modern innovation. Our Director Gillian Hadfield is a leading expert on the need for rethinking our approaches to law and regulation in the twenty-first century. Existing legislative models such as those based on the Fair Information Principles, concepts of de-identification, and the reliance on consent, in addition to the weaknesses identified above, are also especially misaligned with the needs of the modern global digitized world. These approaches are characterized by high levels of ambiguity--such as what constitutes "de-identified" data--which raise legal risks that make data-sharing agreements expensive and hard to reach. They also contribute to the abuse of 'consent' as a means of effectively shifting the power to determine the contours of our emerging data ecosystem to private technology platforms and away from the public. There is a great opportunity to develop new data governance models that can simultaneously improve the democratic quality of data governance and unlock the power of big data and new technologies such as machine learning for public benefit; our Director is at the fore globally in developing these new approaches.

Conclusions

Many members of our research community have active research projects that might be of interest to you in your law reform project. On our Leadership Team alone we are working on a number of projects of potential interest. For example, Peter Loewen and PEARL (Policy, Elections, and Representation Lab) are working in SRI to explore the trade-offs individuals are willing to make between health and COVID related data and community safety, Lisa Austin and David Lie are examining data trust models, and Wendy Wong is researching data rights from the perspective of international human rights law. As University researchers, we have the luxury of taking a broad perspective

[11] Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, Nicolas Papernot, "Machine Unlearning" (https://arxiv.org/abs/1912.03817).

and the long view. However, our Institute is dedicated to building solutions on the basis of careful research and has developed innovative techniques for solutions work. Just prior to the coronavirus lockdowns, we designed and led a very successful initiative focused on addressing privacy obstacles to data access for health care in the context of diabetes. This work played a central role in the effort to stand up the Ontario Health Data Platform to enable the recruitment of big data and machine learning technologies to the fight against COVID-19. If we can be of any further help to you in your law reform project, please let us know.

Sincerely,

Gillian K. Hadfield
Schwartz Reisman Chair in Technology and Society, Professor of Law, and Professor of Strategic Management
Director, Schwartz Reisman Institute for Technology and Society
University of Toronto
Faculty Affiliate, Vector Institute for Artificial Intelligence

Lisa Austin

Lisa M. Austin
Chair in Law and Technology, Professor of Law
Research Lead, Schwartz Reisman Institute for Technology and Society
University of Toronto