WORKING DRAFT DOCUMENT - SPARC
Principles for Using Data Analytics and AI tools

A number of documents that point to principles for the use of AI have been made publicly available in the recent past.
The Berkman Klein Center at Harvard University recently released an analysis of Ethical and Rights-Based approaches to using AI, based on a survey of 36 prominent documents on AI principles[1]. In addition, a Dutch task force, tasked by a number of local academic and research institutions, has prepared a draft set of principles[2] that are still in a period of open consultation, but that has provided the basis for negotiating the agreement on Read & Publish and "academic intelligence" with Elsevier that was recently signed.

While some of the principles overlap, most do not, and for good reason: the Dutch document is designed to guide relationships between academic and research institutions and third party organizations, while the Berkman Klein document surveys policies that are also designed to affect internal documents such as codes of ethics and data policies on the one hand, and legislative programs of governments at the other end of the spectrum.

Looking at the specific needs of academic institutions, we have identified principles that should affect internal documents (i.e. codes of ethics, data policies), principles that should drive relationships with third parties, and principles that affect both (Exhibit 1). In the following page we will provide a brief description of each principle and some examples of how to translate it into concrete steps (such as provisions in internal codes or contractual terms and conditions)

Exhibit 1
A map of AI principles

Internal	Both internal and third-party	Third-party
Areas of application	Strong Privacy Protection	Fairness and non-discrimination
Accountability	Safety and security	Professional responsibility
Human control of technology	Transparency and explainability	Promotion of human values
Governance (internal)	Interoperability	Governance (JVs)
		Ownership of data
		Enduring access

[1] https://cyber.harvard.edu/publication/2020/principled-ai

[2] https://www.vsnu.nl/files/documenten/Nieuwsberichten/Guiding%20Principles%20on%20Management%20of%20Research%20Information%20and%20Data_11May.pdf

PRINCIPLES FOR INTERNAL GOVERNANCE

Areas of application

Identifying which areas of activity of academic institutions should *not* be AI-based is just as important as the actual deployment of AI. Different institutions may differ on what to exclude explicitly, every institution should debate what it wants to exclude, and on what grounds. For example, staff assessment processes may be aided by data collection but also might explicitly exclude using data or algorithms to rank or assess performance. Other activities that could be reasonably excluded altogether are disciplinary actions, staff recruiting and student admissions.

AI can encode the bias of its creators within algorithms in ways that can be difficult to detect and present as "objective" measures[3]. Because of this, it is important that the process of deciding which areas are appropriate for the application of AI be inclusive by design: including regular, meaningful consultation with the full campus community and ensuring the support of those from underrepresented groups who are targets of this bias.

Accountability

Accountability is a broad principle, aimed at ensuring that the "difference" between human and artificial intelligence is bridged as much as possible through mechanisms that allow to "mimic" the accountability safeguards that exist for human decisions. The three overall themes under this category are measures aimed at:

- Designing tools that are verifiable and replicable and that can be evaluated in terms of expected impact
- Using tools that can be held accountable through monitoring bodies and through appeal processes
- Offering redress through remedies for automated decisions and explicit liability.

An example of how this principle could be codified in practical terms is through the design of an identifiable and accessible process to appeal a decision that is made through an automated process. This process could include the right to review the replicability of outcomes and the review of any human intervention after a decision is made by an algorithm, in order to understand whether AI decisions are merely "rubber-stamped" by humans.

Human control of technology

This principle stands for the right of individuals to demand that humans, rather than machines, are the ultimate decision-makers. This principle can include a right to demand human review of any decision that is made through algorithms and the right to opt out altogether without any retribution from automated processes. The application of this principle could also involve the explicit identification of humans who are ultimately

[3] For further analysis, see *Algorithms of Oppression* by Safiya Umoja Noble

responsible for the deployment of AI tools, so that their use (and misuse) cannot be blamed on machines alone.

Governance (internal)

The explicit identification of humans who are ultimately responsible for the deployment of AI tools, academic institutions is just an example of the need to adopt a structured process for the adoption and usage of AI. This process should start with the adoption of a comprehensive set of principles, which should then translate into a set of specific provisions on internal "constitutional" documents, such as Codes of Ethics, Data Policies and Procurement Policies. In addition, proper governance requires the identification of specific individuals and committees (both permanent and temporary) tasked with specific responsibilities of oversight and monitoring of the usage of AI. Whenever possible, representatives of all relevant constituencies (administration, faculty, staff, students) should be involved in the design stages (adoption of principles, definition of specific provisions in internal documents), as well as in the subsequent oversight activities.

PRINCIPLES FOR RELATIONSHIPS WITH THIRD PARTIES

Fairness and equity

Principles that are deemed essential in how academic institutions run themselves must not be violated for the sake of the convenience of using third party tools. Of course, legislation prohibits the most overt discriminatory actions, but AI can pose a risk for discrimination that may be more subtle but is no less real.

The issues of fairness and how to address biases are central for the responsible use of AI. Significant biases can be introduced at each level: by the dataset being analyzed, the mechanism of analysis itself, and the application of that analysis to decision making. Addressing these biases should be a consistent, iterative process. In fact, it is questionable whether bias-free algorithms can even exist. There are a variety of ongoing concrete steps which can be demanded in the relationship with third parties to minimize violations of this principle, such as:

- Reviewing inputs for bias: demanding the use of high quality (i.e. accurate, valid, consistent and appropriate) and audited data sets that have been evaluated for bias
- Reviewing outputs for bias: carefully evaluating the analyses produced by AI tools for outcomes that disadvantage marginalized groups.
- Reviewing implementation for bias: evaluating the application of AI-based tools to avoid the introduction of bias in the use of their analysis.
- Ensuring review by communities negatively impacted by bias: committing to have those who could be disadvantaged by bias review AI-based tools to ensure that they are inclusive, both by design and in practice.

Professional responsibility

Third parties handling data produced by or sold to academic institutions should be held to high standards of professional responsibility. This principle includes a variety of concrete categories:

- Accuracy. This category is largely self-explanatory. It should be noted, however, that accuracy refers both to the quality of the data used, as well as of the algorithms.
- Scientific integrity. This category is a reference to meaningful metrics that should be chosen because they are adequate in quantifying variables that are relevant to measure their intended objective, rather than just because they are easily obtainable. For example, assessing the impact of research or its productivity through the impact factor of journals appears a violation of this principle, as the impact factor was never meant to become a tool for research evaluation.
- Multi-stakeholder collaboration. This concept includes the need for tool developers to consult with regularity and to take into account the impact and implications of their design on both the direct users of the tools and the constituencies that may be ultimately affected . For example, an AI tool that is used for assessing the probability of developing genetic-induced diseases should be developed with adequate input from the life sciences community, as well as from medical practitioners and representatives of patients groups.
- Consideration of long-term effects. Third parties should demonstrably take into account the implications of their work on society, including marginalized communities. For example, a tool that prioritizes funding life sciences research on diseases that affect primarily wealthy countries at the expense of research on diseases that affect primarily countries in the global south would be in violation of the principle of professional responsibility, in addition to other ones.
- Responsible design. Third parties should demonstrate their actual commitment to continuing informing their staff on the broader issues affecting the global community, so that developers can design tools that address broad societal trends, rather than ignore them.

Promotion of human values

Technology development should be for the benefit of human priorities and rights. Therefore, third parties should be in clear observance of human rights, including the well-being of their employees and their working and living conditions. Similarly, technology should promote the well-being of individuals affected by it within academic institutions. Technologies that increase stress, anxiety, or lead individuals to feeling harassed should be banned. In other words, AI systems should not just be safe and in compliance with the law, but also designed to improve the well-being of people using them.

Governance (external)

Whenever possible, third parties operating AI systems with academic institutions should be required to set up oversight boards. The goal of these boards is to create and manage appropriate mechanisms that allow academic institutions to monitor and enforce agreements without escalating them to arbitration or to the judicial system, as well as resolve conflicts. Most important, these oversight boards should be tasked with the ongoing review of compliance of non-contractual elements (for example, like the application of

principles like fairness or promotion of human values, which may prove difficult to insert in contractual obligations).

Ownership of data end enduring access
Academic institutions should maintain ownership of any data and metadata they contribute to third parties. In addition, academic institutions should have perpetual rights to the output of analysis of data and metadata they have contributed, even if the output combines their data and metadata with other sources.

PRINCIPLES APPLICABLE BOTH TO INTERNAL GOVERNANCE AND TO RELATIONSHIPS WITH THIRD PARTIES

Strong privacy protection
The general framework for data protection applicable to Europe (GDPR) represents a good starting point for defining privacy protection. In North America, we would urge academic institutions to adopt many of the GDPR requirements and expand on them. For example, members of academic institutions should:

- Give explicit consent to the collection of data and only be asked to do so after receiving adequate explanation of how it will be used; any changes to the conditions initially agreed upon should also be clearly explained
- Have a right to restrict data usage to a specific list of tasks and activities. Assent should be given explicitly for every activity, rather than through a list providing for blanket acceptance. There should be no negative consequences deriving from the decision to opt out partially or totally.
- Have a right to demand erasure after a predefined period of time, as well to demand rectification of data that is wrong or incomplete or misleading.
- Within applicable laws, should receive notification of any data request from any government or government agency; the data should be handed over only in presence of a legitimate court order.
- Should be immediately notified of any data breach.

Safety and security
This principle requires limited explanation and most data policies of North American academic institutions already cover these issues extensively.

Transparency and explicability
This principle is about ensuring that individuals have an opportunity to understand how technology works, rather than being opaque and complex. The specific categories of actions covered by these policies include

- Using open source software whenever possible, to ensure that algorithms can be independently tested for reliability and to understand biases as well as to allow new entrants and individuals wishing to do so to build applications that build on existing algorithms and data sets.

- If open source software is not available, obtaining rights to independent, third party verification for the purpose of detecting errors and biases whenever open source software is not available
- Translating technical concepts so that individuals can understand how data is used, how answers are generated and what the consequences are.
- Adopting Open Procurement practices . This should be agreed upon in every negotiation, and NDAs (total or partial) should not be accepted. Lack of transparency in contracts, and NDAs in general, have put academic institutions at a disadvantage when negotiating with third party vendors and should be avoided.
- Notifying AI deployment. Individuals should be notified when they are interacting with AI, when data from an activity is being collected for the purpose of successive analysis, and when AI makes decisions on an individual.

Interoperability

This principle is about ensuring that the data analytics environment is decentralized and open to competition. Parties should commit to agreements on standards that support the twin goals of avoiding vendor lock-in and sustaining innovation.