DATA7002:Responsbible Data Science 2021

A/Professor Andrew Crowden
Dr. Slava Vaisman
Dr. Hongzhi Yin
Mr. Hamish MacDonald

Ethics and Data Science: Outline

Week One: Introduction

1. Thinking about Data Science – key questions and definitions
2. Introduction to practical ethics and the nature of moral inquiry and philosophical analysis (dilemmas and conflicts)

Week Two: The theoretical tools of philosophical analysis

3. Approaches to philosophical ethics

Week three: Case study activity

Week Four: Data science governance and regulation

4. What is data and what is information?

5. Collection use and management of data and information

6. Australia's data landscape

Week Five: Decision-making and problem solving in data science case analysis

7. Domain analysis: Research data, Non research data, Algorithm development (machine learning, AI, robotics) and The Practice/s of Data Science.

You are captured in a room. . .

Listen carefully to the problem.

You have to determine the one question that you can ask to either person. Their response will enable you to work out the right path to freedom.

Ethics and Data Science: Outline

Week One: Introduction

1. Thinking about Data Science – key questions and definitions
2. Introduction to practical ethics and the nature of moral inquiry and philosophical analysis (dilemmas and conflicts)

Week Two: The theoretical tools of philosophical analysis

3. Approaches to philosophical ethics

Week three: Case study activity

Week Four: Data science governance and regulation

4. What is data and what is information?

5. Collection use and management of data and information
6. Australia's data landscape

Week Five: Decision-making and problem solving in data science case analysis

7. Domain analysis: Research data, Non research data, Algorithm development (machine learning, AI, robotics) and The Practice/s of Data Science.

4. What is data and what is information?

Data Science

- At its core, data science involves using automated methods to analyze massive amounts of data and to extract knowledge from them. With such automated methods turning up everywhere from genomics to high-energy physics, data science is helping to create new branches of science, and influencing areas of social science and the humanities. The trend is expected to accelerate in the coming years as data from mobile sensors, sophisticated instruments, the web, and more, grows. In academic research, we will see an increasingly large number of traditional disciplines spawning new sub-disciplines with the adjective "computational" or “quantitative” in front of them. In industry, we will see data science transforming everything from healthcare to media. (NYU)

What is Data?

Data refers simply to a collection of material, which can include characters, text, words, numbers, pictures, sound or video. Data may be stored digitally or in hard copy formats, with digitisation enabling data to be copied, stored and transferred rapidly.

The terms ‘data’ and ‘information’ are often used interchangeably. Data can refer to raw data, cleaned data, transformed data, summary data and metadata (data about data). It can also refer to research outputs and outcomes. Likewise, information takes many different forms. Where information is in a form that can identify individuals, protecting their privacy becomes an issue.

‘data’ is intended to refer to bits of information in their raw form, whereas ‘information’ generally refers to data that have been interpreted, analysed or contextualised.

Data

Data is potential information prior to anyone being informed by it

(Jeffery Pomerantz, 2015, p. 26)

QUOTE OF THE DAY (NATURE BRIEFING 31/07/19

"'What does the data say?'

Data doesn't *say* anything. *Humans* say things.

"Information is only as useful as its quality and the skills of the person wielding it, says data scientist Andrea Jones-Rooy. (Quartz)

What is Data?

Data is an imperfect approximation of some aspect of the world at a certain time and place.

It's what results when humans want to know something about something, try to measure it, and then combine those measurements in particular ways.

Andrea Jones-Rooy, Quartz 31/07/19

Metadata

Metadata is a map. Metadata is a means by which the complexity of an object is represented in a simpler form.

(Jeffery Pomerantz 2015, p. 26)

Ethics and Data Science: Outline

Week One: Introduction

1. Thinking about Data Science – key questions and definitions
2. Introduction to practical ethics and the nature of moral inquiry and philosophical analysis (dilemmas and conflicts)

Week Two: The theoretical tools of philosophical analysis

3. Approaches to philosophical ethics

Week three: Case study activity

Week Four: Data science governance and regulation

4. What is data and what is information?
5. Collection use and management of data and information
6. Australia's data landscape

Week Five: Decision-making and problem solving in data science case analysis

7. Domain analysis: Research data, Non research data, Algorithm development (machine learning, AI, robotics) and The Practice/s of Data Science.

5. Collection use and management of data and information

important to acknowledge that . . .

Gathering, understanding, interpreting and making decisions based on collected data is an invaluable tool for, science, business and governments.

(Data Availability and Use, Australian Productivity Commission, March 2017)

there are good reasons for having open data. It speeds research, allowing others to build promptly on results. It improves replicability. It enables scientists to test whether claims in a paper truly reflect the whole data set. It helps them to find incorrect data. And it improves the attribution of credit to the data's originators.

(Nature editorial 12 June, 2017)

Readings

What is data ethics?

Luciano Floridi and Mariarosaria Taddeo

What does it mean to embed ethics in data science? An integrative approach based on microethics and virtues

Louise Bezuidenhout Emanuele Ratti

Data Science case analysis in 4 areas

a) Research **Data**;

b) Non-Research Data
(Collection, Storage and access);

c) **Algorithms** (in machine learning, AI, robotics) and;

d) Data Science **Practices** (responsibilities of data scientists, organisations, data science code of ethics, the characteristics of good data science practice, surveillance)

Ethics and Data Science: Outline

Week One: Introduction

1. Thinking about Data Science – key questions and definitions
2. Introduction to practical ethics and the nature of moral inquiry and philosophical analysis (dilemmas and conflicts)

Week Two: The theoretical tools of philosophical analysis

3. Approaches to philosophical ethics

Week three: Case study activity

Week Four: Data science governance and regulation

4. What is data and what is information?
5. Collection use and management of data and information
6. **Australia's data landscape**

Week Five: Decision-making and problem solving in data science case analysis

7. Domain analysis: Research data, Non research data, Algorithm development (machine learning, AI, robotics) and The Practice/s of Data Science.

6. Australia's Data Landscape

Data Science case analysis in 4 areas

a) Research Data;

b) Non-Research Data
(Collection, Storage and access);

c) Algorithms (in machine learning, AI, robotics) and;

d) Data Science Practices (responsibilities of data scientists, organisations, data science code of ethics, the characteristics of good data science practice, surveillance)

Let's consider what research guidelines tell us about data identifiability . . . .

Data and information may include, but not be limited to:

- what people say in interviews, focus groups, questionnaires/surveys, personal histories and biographies;
- images, audio recordings and other audio-visual materials;
- records generated for administrative purposes (e.g. billing, service provision) or as required by legislation (e.g. disease notification);
- digital information generated directly by the population through their use of mobile devices and the internet;
- physical specimens or artefacts;
- information generated by analysis of existing personal information (from clinical, organizational, social, observational or other sources);
- observations;
- results from experimental testing and investigations; and
- information derived from human biospecimens such as blood, bone, muscle and urine.

Before 2018 in Australia

data may be collected, stored or disclosed in three mutually exclusive forms:

1. **Individually *identifiable* data**, where the identity of a specific individual can reasonably be ascertained. Examples of identifiers include the individual's name, image, date of birth or address;
2. ***re-identifiable* data**, from which identifiers have been removed and replaced by a code, but it remains possible to re-identify a specific individual by, for example, using the code or linking different data sets
3. ***non-identifiable* data**, which have never been labeled with individual identifiers or from which identifiers have been permanently removed, and by means of which no specific individual can be identified (a subset of non-identifiable data are those that can be linked with other data so it can be known that they are about the same data subject, although the person's identity remains unknown reminds us that.

   (*National Statement on Ethical Conduct in Human Research (2007 updated 2015)*

Note: NS public consultation that was due December 21 2016 – current 3.1, 3.2, 3.3 becomes 3.1

Current 3.5 revised updated to 3.5 Human Genomics

- The de-identification of data is a process where data is coded or encrypted. It does not describe data per se. Using the term “de-identified data” to describe how data is stored is inaccurate, misleading and is probably best avoided.

Can you identify and describe an example of:

1. identifiable data
2. re identifiable data
3. non-identifiable data

But do these categories adequately describe data?

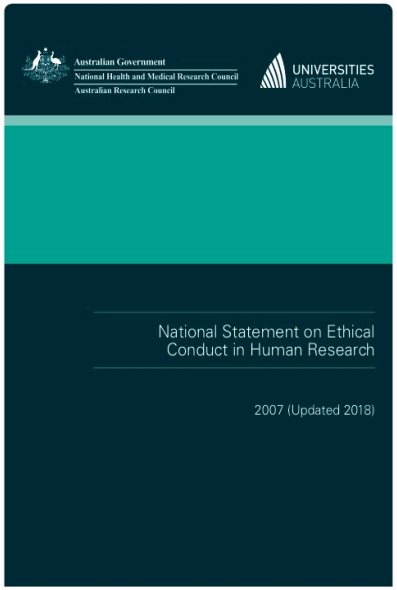
The REVISED NS 2018

The revised National Statement does not use the terms ‘identifiable’, ‘potentially identifiable’, ‘re-identifiable’, ‘non-identifiable’ or ‘de-identified’ as descriptive categories for data or information due to ambiguities in their meanings. Re-identification and de-identification are best understood as processes that change the character of information and are only used with this meaning.

A better way of thinking about data

The identifiability of information is a characteristic that exists on a continuum. This continuum is affected by contextual factors, such as who has access to the information and other potentially related information, and by technical factors that have the potential to convert information that has been collected, used or stored in a form that is intended to protect the anonymity of individuals into information that can identify individuals. Additionally, contextual and technical factors can have a compound effect and can increase the likelihood of re-identifiability and the risk of negative consequences from this in ways that are difficult to fully anticipate and that may increase over time.

AIATSIS Code of Ethics for Aboriginal and Torres Strait Islander Research and the guide.



National Statement 2007 (updated 2018)

NHMRC (National Health and Medical Research Council) (2010). Biobanks Information Paper 2010, NHMRC, Canberra.

Guiding Documents:

- **Australian Code for the Responsible Conduct of Research** :
  - Researchers must comply with ethical principles.
  - Written approval from appropriate ethics committees, must be obtained when required.

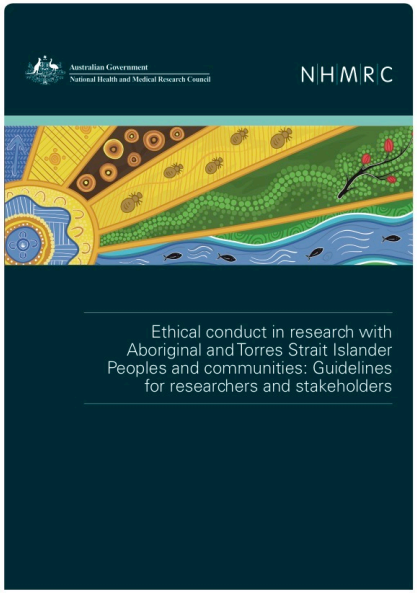
- **National Statement on Ethical Conduct in Human Research**
  - Supports researchers and reviewers to work towards National Human Research Ethics Standards.

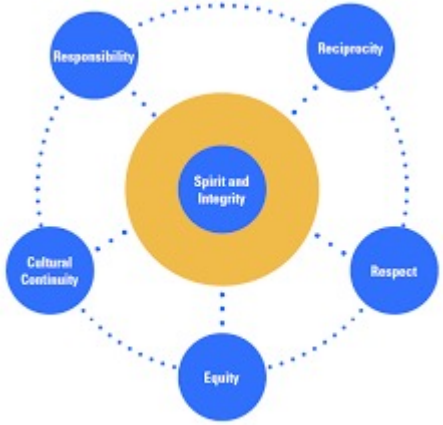
Key Definitions:

- **Research:** investigation undertaken to gain knowledge and understanding or to train researchers.

- **Human research:** research conducted with or about people, their data or their tissue.

The six core values – spirit and integrity, cultural continuity, equity, reciprocity, respect, and responsibility





International/National guidelines for research governance and ethics review

Comparison of international & national statements of research ethics reveals common essentials

International Standards and the National Statement

Social/scientific value	Research merit and integrity
Scientific validity	
Fair subject selection	Justice
Favourable risk benefit ratio	Beneficence
Informed consent	Respect for human beings
Respect for potential & enrolled participants	
Independent review	HREC review

VALUES AND PRINCIPLES

- *Research Merit and Integrity*: scientific standard and scholarly merit, integrity of researchers
- *Justice*: distributive justice - benefits, burdens distributed fairly
- procedural justice - recruitment and ethics reviews involve fair processes
- *Beneficence:* weighing of risks and potential harms, sensitivity to welfare and interests of people, social cultural implications of research
- *Respect for people:* intrinsic value of all people, human autonomy, self determination

a) Research Data

Data Science case analysis in 4 areas

a) Research Data;

b) Non-Research Data
(Collection, Storage and access);

c) Algorithms (in machine learning, AI, robotics) and;

d) Data Science Practices (responsibilities of data scientists, organisations, data science code of ethics, the characteristics of good data science practice, surveillance)

3 major initiatives

A **National Data Commissioner** to support a new data sharing and release framework and oversee the integrity of data sharing and release activities of Commonwealth agencies (2018).

The Australian Competition and Consumer Commission (ACCC) has called for "holistic, dynamic reforms" to address the online dominance of digital behemoths such as Google and Facebook.

AI Principles

Key points re the current situation (identified by the productivity commission)

• New sources of data — as varied as social media sites, smart mobile devices and sensors fitted to physical objects (the 'Internet of Things') — continue to emerge and expand. Digital data, a source of considerable potential value, is being collected ubiquitously.

• The extraordinary capabilities of data analytics and the increasing ability to link previously separate datasets are compounding the usefulness of new data sources, offering important opportunities for better-informed decision making by individuals, businesses and governments, and for research breakthroughs.

• The frameworks and protections for data collection and access, developed prior to sweeping digitisation, require serious re-examination. As one example only, privacy law is neither the only lens, nor even the best, through which to view the use of an asset such as data.

- A shift to viewing data as an opportunity, not necessarily a threat, is a global phenomenon.

• There can be many different competing interests in a particular dataset, including: the subject of the data (such as an individual, who is often also the source); the parties who collect, aggregate and analyse the data; and those who commission these actions. Clarity about these interests is essential to allow Australia to harness the full value of its data.
− The line between what is ‘personal’ data and what is not is blurred both legally and practically. The readiness of individuals to share information about themselves on social media, and other avenues such as loyalty programs, may indicate that social appetite for some types of data use are changing.

A common misperception is that privacy laws — or, indeed, the privacy policies of individual organisations — give individuals ownership over data created by or about them.

- In Australia, no one ‘owns’ data, although copyright law may apply in limited circumstances. Privacy legislation, the primary generic tool offering individuals some control, regulates how personal information is collected, used and disclosed.
- In a world increasingly making use of the data *of others*, the primary unaddressed question should be: for how long will the public — the source of most of this information — trust a structure in which their actual rights are mainly limited to privacy?

An enormous range of information is collected by governments, researchers and businesses about individuals and their activities, institutional and economic structures, and the built and natural environments. However, there is less publication — or controlled sharing — of this information than would help achieve widespread benefits for the community.

• we need guidance on how governments may generate community acceptance of the processes, costs and risks associated with enhanced data use, and to do so where benefits may be most evident.

Productivity Commission Key Recommendations (2017)

Immediate start

Improving data accessibility and quality

- Create **registries** of public and publicly funded data holdings
- **Publish guidance** for custodians on data sharing and release
- Government agencies to commence implementation of **data standards**
- Commence **releasing all** non-sensitive public sector data
- Designate **agency responsible** for introducing data policy reform
- Appoint **National Data Custodian** (NDC) administratively

Improving data usage

- Accredit more **data linkage units**
- **Abolish requirement to destroy** linked data and linkage keys

Data Sharing and Release Act

- Commence exposure draft development

Consumer rights

- Begin industry negotiations on definition of consumer data
- Monitor progress toward target for Comprehensive Credit Reporting (CCR)

Over the next 12 months

- **Draft** Data Sharing and Release Act
  - draft will include an initial list of National Interest Datasets (NIDs)
- **NDC to be tasked** with:
  - developing a process for accrediting release authorities (ARA) and trusted users
  - developing a process for stakeholders to commence nominations for NIDs
  - consultation on draft Act
- Appoint **advisory board** and ethics adviser to the NDC
- Streamline human research **ethics committees processes**
- Priority development of industry **data specification agreements**
- If required, introduce legislation to **mandate** CCR
- **Report on progress** with implementation of new framework

Reform goals

- Publicly funded entities to publish data registries by mid **2018**
- Data Sharing and Release Act to be passed by end **2018**
- From **2019**:
  - NDC to maintain new data framework
  - ARAs to manage National Interest Datasets
  - Data custodians to work with ARAs to improve data use
  - Comprehensive Right enforced by ACCC
- By **2020**, data standards to be operational across the public sector

Government response (2018)

$65 Million Package

- The Government is committed to:

- A **Consumer Data Right** as a new competition and consumer measure to allow consumers to harness and have greater control over their data.
- A **National Data Commissioner** to support a new data sharing and release framework and oversee the integrity of data sharing and release activities of Commonwealth agencies.
- A **legislative package** will streamline data sharing and release, subject to strict data privacy and confidentiality provisions.

Link to data commissioner

https://www.datacommissioner.gov.au/

- The Australian Competition and Consumer Commission (ACCC) has called for "holistic, dynamic reforms" to address the online dominance of digital behemoths such as Google and Facebook.

https://www.accc.gov.au/focus-areas/digital-platforms

Australian Gov and AI Principles

https://www.industry.gov.au/

Link to UQ data

https://data.uq.edu.au/

1) Research and 2) non-research data

We can lean much about data ethics from case examples and individual lived experiences. . .

Clinical Practice/Research

Genomics and precision medicine

Diseases with a genetic cause

- **Monogenic** – mutation of a single gene, three types, *autosomal dominant* (Huntington's) – mutation in one copy of a gene is inherited from either parent, *autosomal recessive* (Cystic Fibrosis) –mutation in two copies of a gene- affected child inherits one copy from each parent, *sex linked*, are traceable not to 22 autosomes but to the sex chromosomes X and Y, most are X-linked meaning 1 in 2 of male children of mothers who carry the mutation will be affected, half the daughter will be unaffected carriers (Duchene muscular dystrophy and hemophilia – high penetrance – close to 100% chance of getting the disease)
- **Polygenic** –mutation of several or more genes (maybe some forms of Ca)
- **Multi-factoral** –the interaction of several or more genes and the environment (most diseases)

Only a small percentage of diseases are caused by a mutation in a single gene

Huntington disease (HD) is a neurological degenerative disease that has an onset in most people between the ages of 30 and 50. There is no cure for this condition and it is progressive. Symptoms include deterioration in movement, cognition and generalised functioning. Death usually results from respiratory illness. HD is an inherited condition. A child of an affected person has a 50% chance of inheriting the faulty gene that causes the condition. Genetic predictive testing is now available for persons over the age of 18 who have an affected parent or relative which will tell them in almost all cases whether they will develop the disease at some stage in their life. Worldwide, of those eligible for the test, only around 15% of people have taken up the option of testing.

Kunmanara is a 25 year old man whose grandfather died some 10 years ago from Huntington disease.

His mother has therefore a 50% chance of developing HD. She decided to have the genetic test and has been shown to have the faulty gene. She will definitely develop HD at some time.

He now has a 50% chance of developing HD. He is an air traffic controller. He loves his job and he feels he could perform his duties most adequately for many years, irrespective of whether he carries the faulty gene for HD or not. His employer is unaware of his family history.

Should Kunmanara have the test?

Afterwards Kunmanara regrets his decision. He rings the health service where the test was done and requests that the test data be destroyed. He does not want to know the result. The case is further complicate by the fact that his HD test result was positive – he will certainly develop HD during his life. When he visited the clinic, he consented to

1. the test.
2. being a participant and contributing his data to a local research project.
3. being a participant and contributing his data to an international research project and
4. allowing his data and information to be included on a international databank. This information and other data had already been added to a global data repository of HD, a [global] hybrid genomic data infrastructure (genomic biobank) by a researcher/data scientist.

To complicate things, the Health service treating team, the GP, the Clinical Geneticist, the Genetic counsellor, the data scientist/researcher and nurses have different views about how to progress.

They refer the case to the Clinical Ethics Committee (CEC), the Human Research Ethics Committee (HREC) and a specialist HREC for advice.

Kunmanara decides to have the test and is seen by a GP (consent for the test), a clinical geneticist (who gets an additional consent for Kunmanara's clinical and other information to be added to a HD international database (approved by a local HREC) and a local researcher who obtains consent is supervising a PhD on Nurse/Genetic Counsellor (who explains all consents).

…and Kunmanara is a Pitjantjatjara man, originally from Amata in the far northwest of South Australia.

The case is also referred to a specialist HREC with particular expertise in targeted Aboriginal and Torres Strait Islander research for advice.

To know or not to know? When is the right time to decide to have predictive/pre symptomatic testing?

Do employers in industries involving public safety have the right to demand family health

history information? In cases where genetic predictive testing is available for conditions

that may impact on public safety, do employers have a right to predictive testing

information about an individual whose current health status is excellent?

Who actually ‘owns’ this information and who should decide who can access it?

What if the situation was reversed and Kunmanara wanted testing but his mother had refused?

What responsibility is there to offer testing to an individual when the result may indirectly

reveal the genetic status of a relative (if Kunamanara carries the HD gene fault then he must have

inherited from his mother)?

Are there implications for his reproductive choices?

What are the ethical obligations of data scientist/clinician/ researcher role/s? Is clinical consent different from research and database consent? Can the data be kept without consent for greater good?

Ethics and Data Science: Outline

Week One: Introduction

1. Thinking about Data Science – key questions and definitions
2. Introduction to practical ethics and the nature of moral inquiry and philosophical analysis (dilemmas and conflicts)

Week Two: The theoretical tools of philosophical analysis

3. Approaches to philosophical ethics

Week three: Case study activity

Week Four: Data science governance and regulation

4. What is data and what is information?
5. Collection use and management of data and information
6. Australia's data landscape

Week Five: Decision-making and problem solving in data science case analysis

7. Domain analysis: Research data, Non research data, Algorithm development (machine learning, AI, robotics) and The Practice/s of Data Science.

Tutorial Readings

What is data ethics?

Luciano Floridi and Mariarosaria Taddeo

What does it mean to embed ethics in data science? An integrative approach based on microethics and virtues

Louise Bezuidenhout Emanuele Ratti