

### Introduction to AI & AI Ethics (Syllabus Outline Document)

(**</>**) AI6101

n Dr. Melvin Chen



(m) Nanyang Technological University

### 👘 Instructor Information

Name Dr. Melvin Chen

Office HSS-03-91 (Humanities & Social Sciences Building, Level 3)

Telephone 65927935

E-mail (preferred) melvinchen@ntu.edu.sg

Consultation Hours 2-4 pm on Thursdays (please e-mail in advance) Personal Website http://thesingaporeanphilosopher.com/

# (d) Meeting Times & Venues

Monday LT 3 / Zoom (hybrid) 6.30-9.30 pm

### Course Details

### NTU Learn

This course will rely heavily on the NTU Learn portal. NTU Learn link: https://ntulearn.ntu.edu.sg Announcements. Please check your e-mail and NTU Learn regularly Readings. Weekly readings to be downloaded from NTU Learn Course Slides. Slides to be uploaded onto NTU Learn after each lecture

### **%** Grade Assessment

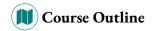
Reading Tasks (Weeks 8 & 9)  $(2 \times 0.05 \times 40)\%$ 

= 4%

Take-home Essay  $(0.6 \times 25)\%$ 

= 15%

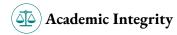
TOTAL 19%



Topic	Questions		
The Philosophical Foundations of AI Ethics (Part 1)	What is the traditional fact-value distinction in episte-		
	mology?		
	How does the value-neutrality thesis about technology		
	work?		
	What are the value alignment problem and the AI con-		
	trol problem?		
	Can you identify any basic AI drives?		
	What are some possible responses to the value alignment problem?		
	Can we determine the correct moral standard and design		
	machines in accordance with that standard?		
	How do we distinguish between consequentialism and		
	deontology as candidate moral standards?		
	What is inverse reinforcement learning and how might		
	it be employed with respect to the value alignment prob-		
	lem?		
	How might we identify fair principles for AI alignment?		
	What is superintelligent AI and how might its arrival		
	help us on the value alignment front?		
The Philosophical Foundations of AI Ethics (Part 2)	What are the existential risks of superintelligent AI?		
	What additional ethical considerations do AI systems pose?		
	Should we recognize the rights of robots and the moral		
	status of AI systems?		
	What are the normative implications of jobs becoming		
	automated?		
	What are some undesirable ends to which AI systems		
	have been deployed?		
	Might AI systems result in the end of the human		
	species?		
	Could we introduce safeguards to deal with AI-relevant		
	ethical considerations?		
	How do we build wisdom and friendliness into our AI		
	systems?		
	What are some principles, declarations, and frameworks		
	that have been devised with respect to AI systems?		
	How might these principles, declarations, and frame-		
	works offer normative guidance?		

## Required Textbooks & Readings

All weekly readings are to be downloaded from NTU Learn. Please note that these readings are to be used solely for educational purposes in connection with curriculum-based learning.



It is the student's responsibility to understand NTU's Academic Integrity Policy. All students should familiarize themselves with the 'Guide to Academic Integrity'

Link: http://www.ntu.edu.sg/ai/ForEveryone/Pages/AGuidetoAcademicIntegrity.aspx

### Course Schedule

Week	Lecture	Date of Class	Topic	Readings	Test, Essay
		Meetings	_		Deadline, or
		_			Other Remarks
Week 8	Lecture 1	4 Oct 2021	The Philosophi-	Bentham	Week 8 Reading
			cal Foundations	(1789), Ross	Tasks (2%)
			of AI Ethics	(1930)	
			(Part 1)		
Week 9	Lecture 2	11 Oct 2021	The Philosophi-	Anscombe	Week 9 Reading
			cal Foundations	(1958), Rus-	Tasks (2%)
			of AI Ethics	sell & Norvig	
			(Part 2)	(2010)	
Week 13					Take-home
					Essay Deadline
					(11.59 pm on 14
					Nov) (15%)

# Reading List

### REQUIRED READINGS (MANDATORY):

#### Lecture 1

Bentham, Jeremy. 1789. An Introduction to the Principles of Morals & Legislation, London: T. Payne & Son, Chap. I & Chap. IV

Ross, W. D. 1930. 'What Makes Right Acts Right,' in *The Right & the Good*, ed. Philip Stratton-Lake, Oxford: Clarendon Press, Chap. II, pp. 16-47

#### Lecture 2

Anscombe, G. E. M. 1958. 'Modern Moral Philosophy,' in *Philosophy*, Vol. 33 No. 124, pp. 1-19
Russell, Stuart & Peter Norvig. 2010. 'The Ethics & Risks of Developing Artificial Intelligence,' in *Artificial Intelligence: A Modern Approach*, 3<sup>rd</sup> ed., Prentice Hall, pp. 1034-40

#### SUPPLEMENTARY READINGS (OPTIONAL):

#### Metaethics

Sumner, L. W. 1967. 'Normative Ethics & Metaethics,' in Ethics, Vol. 77 No. 2, pp. 95-106

#### **Normative Theory**

Aristotle. 1906 [c. 350 B.C.E.]. 'Moral Virtue,' in *Nicomachean Ethics*, trans. F. H. Peters, 10<sup>th</sup> ed., London: Kegan Paul, Trench, Trübner, & Co., Book II, pp. 34-57

Kant, Immanuel. 2005 [1785]. 'Passage from Popular Moral Philosophy to a Metaphysic of Morals,' in *The Moral Law*.

Groundwork of the Metaphysics of Morals, trans. H. J. Paton, Abingdon, Oxon: Routledge, Chap. II, pp. 79-126 – selected portion

#### **AI Ethics**

Omohundro, Stephen. 2018. 'The Basic AI Drives,' in *Artificial Intelligence Safety & Security*, ed. Roman V. Yampolskiy, 1<sup>st</sup> ed., New York: Chapman & Hall/CRC, Chap. III, pp. 47-56

Personal Data Protection Commission Singapore. 2020. *Model AI Governance Framework*. 2<sup>nd</sup> ed. Retrieved 7 Dec 2020 from: <a href="https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf">https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf</a>



# Appendix 1: Assessment Criteria for Take-home Essay Question Essay Response

Grade/Numerical Score	Criteria
	Clarity and distinct originality of thought, with clear link to major topics of the
A to A+ (80-100%)	primary readings
	Compelling use of persuasive and effective argument in every paragraph to support
	claims
	Excellent use of language, with no grammatical errors
	Consistent demonstration of close reading of primary readings and detailed and in-
	depth analysis of the relevant theoretical concepts
	Ability to introduce, review and engage critically with secondary readings (where
	relevant)
	Clarity of thought, with clear link to major topics of the primary readings
	Convincing use of persuasive and effective argument in most paragraphs to support
A- (75-79%)	claims
	Good use of language, with a few grammatical errors
	Some demonstration of close reading of primary readings and detailed and in-depth
	analysis of the relevant theoretical concepts
	Ability to introduce, review and engage critically with secondary readings (where
	relevant)
	Some discernible link between thesis and major topics of the primary readings
	Convincing use of persuasive and effective argument in some paragraphs to support
B to B+ (65-74%)	claims
	Average use of language, with a number of grammatical errors
	Close reading of primary readings and detailed and general analysis of the relevant
	theoretical concepts
	Ability to introduce and review secondary readings (where relevant)
	Almost indiscernible link between thesis and major topics of the primary readings
B- to C+ (55-64%)	Unconvincing and ineffective use of argument
D to C1 (33 04/0)	Average use of language, with serious grammatical errors that threaten clarity of
	expression
	Summarization of primary readings and description of theoretical concepts
	Ability to introduce and review secondary readings (where relevant)
C to D (45-54%)	Clear absence of link between thesis and major topics of the primary readings
	Complete absence of argument or the use of incoherent or invalid argument to sup-
	port claims
	Poor use of language, with serious grammatical errors that threaten clarity of ex-
	pression
	Summarization of primary readings and misinterpretation of theoretical concepts
	Introduction of irrelevant secondary sources
F (0-44)	Failure to submit