HUMANISING MACHINE INTELLIGENCE

Machine intelligence will dramatically alter decision-making in every field of human endeavour, from the mundane to the epochal. But however complex the machine, what you get out is a function of what you put in. New decision technologies will carry the imprint of the society that creates them. Our grand challenge is to design intelligent machines that reflect our best selves; that make decisions that are fair, just, and compassionate. Our grand challenge is to ***humanise machine intelligence****.*

1. IMPACT AND ALIGNMENT WITH ANU STRATEGIC PLAN

People are notoriously bad at making decisions. We prioritise the short term. We overestimate our abilities. We misunderstand probabilities. We are riven with bias. Machine Intelligence (MI) is a technology for improving human decision-making. Whether through the use of machine-learning-driven data analytics to support human operators, or through wholly autonomous systems, MI promises to overcome our biases and cognitive defects, and help us make decisions that are not only more efficient, but also fairer and more just.

To date, however, this promise has emphatically not been realised. Think of 'robodebt', or the replacement of cash bail by algorithmic risk assessment in California, or early failures of self-driving cars, or the campaign against autonomous weapons. There is widespread fear—recently voiced by Nobel Laureate Joseph Stiglitz—that MI will exacerbate, rather than improve, economic and social inequality, and increase the risk of unjust harm on the roads or at war.

What has caused this moral failure? The answer is simple. MI is a decision technology. At its heart is mathematical decision theory. To date, there has been no successful incorporation of *moral reasons* into decision theory. MI no doubt offers efficiency gains. But if the decision theory being optimised is amoral, this simply means more efficient injustice.

The solution? One approach is to retrofit ethics to MI through regulation. Though crucial, the costs of trial and error are too high to rely on that alone. We also need to **design** ethical MI. We need to *humanise* machine intelligence. And let's be clear. **The alternative to ethical MI is not a world without MI. It is a world with unethical MI. Humanising Machine Intelligence is more than a grand challenge: it is a necessity.**

Central to the ANU strategic plan is **research excellence** for **social and economic benefit**. The downside risks of unethical MI are huge. The upside potential of ethical MI is equally great. Humanising MI is a necessary step in mitigating those risks, and realising that potential. This not only enables the creation of tremendous value, it also furthers the goal of **achieving equity** in society. HMI will also **foster collegiality across campus and beyond**, through interdisciplinary engagement, and strong connections with industry and government.

SUCCESS IN 5 YEARS

- Identify the key social risks and opportunities associated with adoption of MI.
- Develop theoretical foundations for incorporating moral values into MI.
- Build code for adoption in autonomous systems and machine-learning-based decision support.
- Build a self-sustaining research institute leading the world in humanising machine intelligence.

REALISTIC PROPOSED IMPACTS

- **Increased public understanding** of risks and opportunities of MI. Feasible because of social science expertise, and experience conveying social research to civil society and government.
- **Major theoretical advances** in decision theory, ethics, machine learning. Feasible because building on existing research excellence.
- **Uptake of algorithms** in government and industry. Feasible because of our track record of major advances in MI with national and global impact.

BUILDING ANU PROFILE

- Uniquely combining world-leading experts in social sciences, philosophy and computer science in a unified approach to **formulate and solve the design problem of humanising MI**.
- As the first project to unify superb decision theorists based across ANU, HMI will establish ANU as one of **world's best centres for decision theory**.
- ANU will become global leader in **philosophy of machine intelligence**.

3. TRANSFORMATIVE RESEARCH

DISCOVERY: WHAT ARE THE RISKS AND OPPORTUNITIES OF WIDESPREAD ADOPTION OF MI?

All technologies are developed to serve a purpose. Advancing MI is not a goal in itself—and adopting MI is no guarantee of improving decision-making. To formulate our design problem we must therefore have in mind how MI will be used. We must explore the existing and inevitable uses of MI, but also those that are as yet uncertain—both malicious and well-intentioned. Research targets would likely include:

MACHINE LEARNING AND DATA ANALYTICS

Cambridge Analytica, 'Robodebt', social media, use of ML in legal risk assessment for bail, parole etc., medical diagnosis, government resource allocation decisions (e.g. welfare, housing, immigration, visas).

AUTONOMOUS SYSTEMS

Trading algorithms, supply chain platforms, self-driving cars, energy allocation, triage systems, rescue robots, companion robots, autonomous defence systems, general AI.

FOUNDATIONS: HOW CAN WE FORMULATE A MORAL CODE FOR MI?

CAN WE FORMALISE MORALITY FOR MACHINES?

Pessimists think machine morality is an oxymoron—perhaps because morality depends on affect; or because it is too complex to be formalised; or formalisation implies crude cost-benefit analysis.

Even optimists often leave a gap between principle and practice—e.g. a recent German government report on self-driving cars identifies attractive ethical principles, but says nothing about implementation.

We are **realistic optimists**. We have shown how to incorporate a range of values into classical decision theory, the mathematical architecture for MI. Our grand challenge will be to generalise this approach:

- Fit a **broader range** of moral theories;
- Develop ***sequential*** **moral decision theory**;
- Account for moral **uncertainty/disagreement**.

WHICH VALUES SHOULD WE FORMALISE?

Given power relations and competing ideologies, can we meaningfully talk of 'our' values?

Yes! There are **'Settled Norms'**—especially in political communities. We must identify these, and operationalise them for probabilistic decision-making.

'Unsettled Norms' can be ***uncertain*** or ***contested***.

- Uncertainty can be addressed mathematically; we will extend this to moral uncertainty.
- Where disagreement obtains, we will provide space within MI systems for reasonable moral judgements, whatever they are—leaving precisification and calibration to end-users.

We will also consider **'design norms'**, e.g. ensuring that MI decision-making is transparent to and interpretable by humans.

DESIGN: HOW CAN WE OPERATIONALISE MORALITY FOR MI?

The next task is to operationalise our answers to those theoretical questions, through case studies that provide **a 'proof of concept' that machine intelligence can be humanised**. Because of our collective expertise in decision theory, our theoretical solutions will already be expressed in the language that underpins MI. We will focus on use-cases falling into two camps: (1) machine-learning-driven decision support and (2) autonomous systems. This technology is fast-moving, and to ensure maximum uptake we will identify core case-studies in the first year, in consultation with industry partners. However, these are examples of what we might pursue:

- Develop design parameters for building fairness concerns into machine-learning algorithms for legal risk-assessment in bail and parole decisions, possible partners include federal/local govt.
- Develop and operationalise ethical decision theory for self-driving vehicles, in particular managing tradeoff between efficiency and risk to others, possible partners include WA mining industry.

Through focus on case studies we will develop a design methodology that can be exported to diverse applications, including general AI. **We will develop (1) systems that can autonomously make moral decisions, and (2) risk-assessment frameworks for MI that elevate high-risk decisions to a 'human in the loop'.**

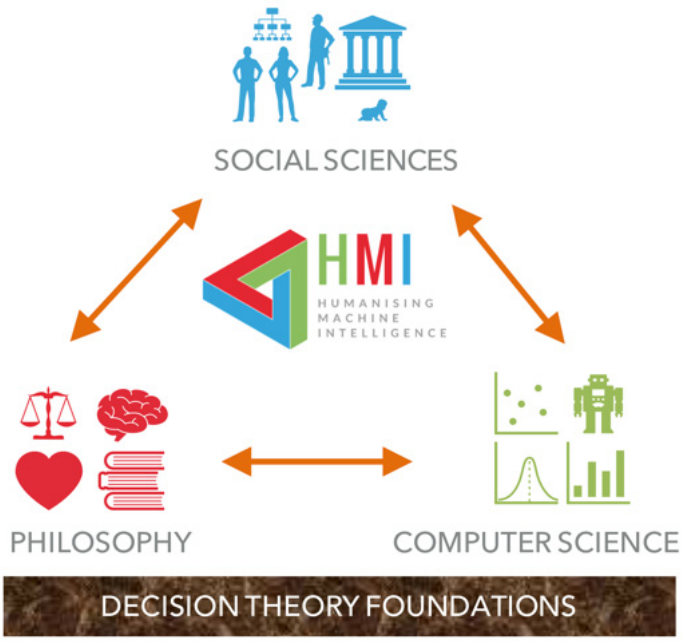
PIONEERING THEORETICAL AND/OR METHODOLOGICAL PARADIGM SHIFTS

- Unified empirical, theoretical, applied approach; not just interpreting the world—changing it.
- Uniting mathematical decision theory across computer science, economics, and philosophy.
- Posing new problems in philosophy and social sciences about complexity, computability, accountability, and justification.
- World-first sequential moral decision theory.

4. APPROACH

No discipline alone can humanise machine intelligence. But, in combination, the strengths of the ANU are uniquely well-suited to the task. Each of our research areas is naturally led by one of our discipline areas—**discovery** by social scientists; **foundations** by philosophers; and **design** by computer scientists. But each topic will be enriched by the insights of the other disciplines: **this is a fundamentally collaborative project**.

- **Discovery**: economics (Meneghel), sociology (Davis) and political science (Erskine) are crucial here, but our team's expertise in the philosophy of psychology (Klein) will also be vital to understanding the human impacts of MI systems, and the experience of our computer scientists (e.g. Xie, Williamson, Thiébaux) in developing existing applications of MI will also be crucial.
- **Foundations**: led by our philosophers, Klein, Lazar and Steele, this area will benefit from social and political theory insights from Erskine and Davis, from Meneghel's ground-breaking work in decision theory, and from the decision-theoretic expertise of Hutter, Thiébaux, Xie and Williamson.
- **Design**: led by Hutter, Thiébaux, Xie, Williamson, but this operationalises the insights of discovery and foundations, so will draw on all the work done so far—and inform it, as design challenges shed light on foundations and discovery.



The three disciplinary bases on which we will build to humanise machine intelligence, on top of the foundation of decision theory. Interactions will occur between all three, and the ECRs we will hire will "live on the edges".

PATHS TO IMPACT

- **Academia**: Global, long-term impact through: publishing world-class research in best journals; hosting ground-breaking seminar series and workshops fostering international collaboration; hosting world's best scholars; partnering with the leading research institutes. Already have positive interest from CFI, AINow, APL, MIT Media Lab and others.
- **Industry and Government**: We have well-established relationships with major government and industry partners. Already received warm initial responses from Chief Scientist at DeepMind, Google Head of AI Public Policy for Asia Pacific, ElementAI, DFAT and others. Other likely partners include Snap, Microsoft, AFOSR, Facebook, Amazon etc. We will choose lead partners and share research through workshops and reports distilling our basic research on discovery/foundations/design. Mnemosyne project will also be crucial link.
- **Civil Society**: Partnership with public outreach organisations with which our team has existing links, e.g. Future of Life, Sloan and Moore foundations, Open Philanthropy, and our own 3Ai. Public events, active social media, commentary. Co-production of podcast series with *Hi-Phi Nation* to reach global audience.

MILESTONES ([N] = COMPLETED N YEARS AFTER COMMENCEMENT)

- [1,2,3,4,5,6,7] Fortnightly seminar series; publication of basic research in high-impact journals; additional funding sought from government, industry, civil society; annual advisory board meetings.
- [1] Mission statement published; first case studies decided; first postdocs/HDRs in place; agreements with partners in industry, govt, academia.
- [1,2,3]2 workshops/yr:**Discovery**, **Foundations.**
- [2] Podcast series with Hi-Phi Nation launched.
- [3] Publication and promotion of **Discovery** report, workshops with govt/industry partners.
- [4] 2 workshops each on **Foundations, Design.**
- [4] Publication and promotion of **Foundations** report, workshops with govt/industry partners.
- [5] Match GC funds with external awards for Y6.
- [5] Research workshops on **Design**; publication/promotion of **Design** report, workshops with govt/industry partners.
- [6] Use of our code in government and industry; new partnerships to develop new research programmes around HMI theme.
- [7] Establish self-sustaining HMI centre.

RISKS AND MITIGATIONS

- Moral complexity proves intractable: *aim then becomes better approximating morality.*
- No uptake of results: *deliberate working back from design problem—results usable by design. Member of exec explicitly tasked with overseeing impact.*
- Key person risk: *each discipline has multiple representatives; team has several experienced leaders.*
- Overtaken by academic competition: *unique approach, international reputations, head-start.*
- Overtaken by industry: *focus on basic research for long-term impact—at which ANU excels.*
- Inability to attract talent: *use existing disciplinary strengths and topicality as key attractors.*
- Cannot distil theory to design guidance: *MI systems built on decision theory; building moral reasons into this framework gives a path forward.*
- Technological advances render existing MI obsolete: *cutting-edge team will adapt fast to technological leaps; new forms of MI will need ethical design.*

5. ANU COMPETITIVE ADVANTAGE

The Chief Scientist believes **Australia can lead the world in ethical MI**. We agree. But to do so, we must invest in work that carries the insights of the social sciences and humanities over to computer science. Computer scientists cannot create ethical MI alone. But discourses on MI ethics are useless if they cannot be operationalised. **No existing research group combines our calibre of empirical expertise, theoretical insight, and engineering aptitude, all focused on a common goal: to design humanised machine intelligence.** There are many excellent research institutes working on related problems. We hope to collaborate with most of them, and have already had warm exchanges with several. But each of them differs from us in crucial ways:

- Some grounded in computer science, with other disciplines peripheral—Alan Turing Inst., Machine Intelligence Research Inst. (Berkeley), Centre for Human-Centric AI (Berkeley), OpenAI.
- Some target law/regulation, not design—AINow (NYU), Ethics + Governance of AI (MIT/Harvard).
- Some focus on further future, not near- to mid-term—Future of Humanity Institute (Oxford), Centre for Study of Existential Risk (Cambridge).
- Some focus on defence and security—Trusted Autonomous Systems, APL (Johns Hopkins), Cybersecurity Institute (ANU).

The **Leverhulme Centre for the Future of Intelligence** (Cambridge) is the world's top research institute in this area. It will be a close collaborator; we are already planning two joint workshops for 2019. It encompasses many disciplines, but researchers work on separate projects, not on a common task as we propose to do.

Another collaborator: **3A Institute (ANU)**. HMI designed in consultation with Prof Bell (who will be on advisory board) to ensure **HMI complements 3Ai**. Particular synergies with respect to **Discovery** and **Engagement**. Differences: **3Ai focuses on *managing* cyber-physical systems; HMI on *designing* ethical MI**. 3Ai: aims to create new applied science, project leader's roots in anthropology; HMI: cross-disciplinary collaboration of research leaders from computer science, philosophy, social sciences. 3Ai: 'Innovation Institute' targeting new forms of education/engagement. HMI: basic research institute, aiming at publications in high-impact journals, competitive grant funding etc.

ANU COMPETITIVE ADVANTAGE AND TRACK RECORD

Our computer scientists are at the forefront of machine intelligence research, have won major awards, and millions of dollars in grants from govt and industry. ANU Philosophy is top 5 worldwide for decision theory, top 2 for political philosophy (www.philosophicalgourmet.com, www.bit.ly/PGRANU). We are 13th in the world for Sociology, 8th for Politics and International Studies (QS 2018). Few universities can boast the same suite of skills. **Nowhere else has a team combining comparable expertise with the same unity of purpose**.

6. TEAM AND GOVERNANCE

PROJECT LEADER (0.4FTE)

- Seth Lazar, head of School of Philosophy. During his career, no other philosopher has published more frequently in the two leading moral/political philosophy journals.
- Part of every philosophy hire since joining ANU. HoS since 2017. Steering Philosophy through regeneration, rise in international rankings. Devised Centre for Philosophy of the Sciences.
- Has breadth needed to integrate whole project: is team's leading ethicist; trained in political science; has developed a 'deontological decision theory'—building duties into decision theory.
- ECR (just), but has already worked with audiences outside academia, e.g. US military. Active interest in advancing public engagement.
- Will benefit from oversight of an **advisory board**.

EXECUTIVE (ALL 0.3FTE)

- **Toni Erskine**, **Colin Klein**, **Sylvie Thiébaux**, **Bob Williamson**. Collective experience leading organisations with $5m to $100m pa budgets.
- **Team project**. All of our research questions require contributions from all team members, and **executive members will work on all questions**.
- **Erskine will oversee Discovery**. One of Australia's most respected politics/international relations scholars. Working with Google on AI; associate fellow of the CFI. Her expertise in political theory crucial for navigating moral disagreement.
- **Klein will oversee Foundations**. A leading philosopher of psychology and cognitive science (among other areas), Klein publishes in leading journals, and has track record of collaborative cross-disciplinary research that commands high levels of public interest (e.g. bee consciousness).
- **Thiébaux will oversee Design**. An experienced leader, Thiébaux has won numerous awards for work on AI theory and practice, including deploying optimising AI in award-winning Bruny Island Battery trial, and is co-editor of the leading journal in the field. She has many ARC and industry grants.

- **Williamson will direct outreach to government and industry**. Fellow of Australian Academy of Science, has led multiple major research projects, with significant govt/industry support, including helping found NICTA (now Data61). Nobody better placed to ensure uptake of our research.

TEAM (ALL 0.2FTE)

- **Jenny Davis**, ECR sociologist editing prominent *Cyborgology* blog. Highly-published and -cited per career stage. ANU Futures award winner. Primary contributions: **Discovery** and **Design.**
- **Marcus Hutter**, world's leading authority on universal AI. Researching AI, machine learning, philosophy, decision theory. Major grants from industry, govt, civil society. **Foundations** and **Design**.
- **Idione Meneghel**, economist specialising in decision and game theory. Expert on uncertainty, publishing in leading economic theory journals. **Discovery**, **Foundations**, **Design**.
- **Katie Steele**, leading decision theorist with influential work on uncertainty and on moral decision theory. **Foundations** and **Design**.
- **Lexing Xie**, highly-decorated expert in machine learning and social media. 2018 Computer Science Association award. **Discovery** and **design.**

CITATION DATA

Computer science, sociology, political science are high-citation disciplines. Economic theory and philosophy are not. Adjusting for career stage and discipline, our team is highly cited. Total/h-index.

COMPUTER SCIENCE AND SOCIAL SCIENCES

- Davis 363/9; Erskine 1364/16; Hutter 4705/31; Thiébaux 2488/28; Williamson 17311/47; Xie 4681/31.

ECONOMICS AND PHILOSOPHY

- Klein 799/14; Lazar 481/12; Meneghel 72/3; Steele 438/10.

7. FUNDING

BUDGET

This indicative budget covers expenditure types and prospective annual totals. We are bidding for $10m over seven years. By end Y5 we aim to match GC funds with external awards. By end Y7 we aim to be self-sustaining. We will front-load GC funds to generate momentum to win other backing.

PERSONNEL

- 6 (Y1-Y2) to 10 (Y3-Y7) postdocs: up to $140,000pa each.
- 1 (Y1-Y2) to 2 (Y3-Y7) RA and administrator: up to $100,000pa each.
- 4 (Y1-Y2) to 6 (Y3-Y7) PhD scholarships: $30,000pa each.
- Teaching relief (especially ECR): up to $70,000pa.

ENGAGEMENT AND OUTREACH

- Marketing/social media/visual communications/publishing/PR: up to $50,000pa.
- Production of public outreach resources including podcast series: up to $20,000pa.

RESEARCH ACTIVITIES

- Fortnightly seminar series and 2-4 workshops/year: up to $80,000pa.
- 10+ 4-8wk visiting fellows: up to $100,000pa.
- Travel (especially ECRs): up to $100,000pa.
- Open access publishing: up to $30,000pa.
- Equipment, computation access: up to $40,000pa.
- Fieldwork and experiments: up to $50,000pa.

INDICATIVE ANNUAL SPEND

- Total spend. Y1: $1m. Y2: $1.5m. Y3: $2m. Y4: $3m. Y5: $3m. Y6: $2.5m. Y7: $2.5m.
- Split GC/External. Y1: 1m/0m. Y2: 1.25m/0.25m. Y3: 1.5m/0.5m. Y4: 2.25m/0.75m. Y5: 2m/1m. Y6: 1.25m/1.25m. Y7: 1m/1.5m.

ATTRACTING OUTSIDE FUNDING

GOVERNMENT

- Will apply for ARC grants (Discovery, DECRA, Future Fellowship, Linkage, Centre of Excellence).
- Will build on established links with CRC scheme, Data61, US Air Force Office of Scientific Research.

INDUSTRY

- Strong track record of funding from global tech.
- Potential for Industrial Training Transformation Centre funding for HDR/postdocs.

CIVIL SOCIETY

- Comparable research institutes have raised substantial philanthropic funding (MIRI, Future of Humanity Institute, CFI).
- Early discussions and track record suggest we are well-placed to pursue such funds, from e.g. Future of Life, Moore foundation, Sloan Foundation, Templeton World Charity Foundation.

APPENDIX 1: FURTHER DETAILS ON TEAM

OVERVIEW

Each member of this team is, per career stage, at the top of their field. And yet the team is still more than the sum of its parts. We have the breadth needed to lead each stage of our enquiry, while maintaining a unity of purpose that will ensure a genuine research community. Text in red = evidence for experience mentoring.

IDIONE MENEGHEL (PHD+5)

Associate editor of *Journal of Mathematical Economics*. Published in top journals in economic theory. Works on reasoning under uncertainty, also game theory. Recently visiting lecturer at Yale. Has supervised university medallist honours student, and PhD student placed on postdoc at Humboldt.

JENNY DAVIS (PHD+6)

Author of forthcoming book with MIT on social impacts of technological artefacts. Qualitative, theoretical, and experimental research into human behaviour. Co-editor of Cyborgology blog. Published in leading international sociology journals. ANU Futures award holder. Already on five HDR committees and has an active cohort of honours students.

SETH LAZAR (PHD+8)

Highly published in premier philosophy journals. Monograph on moral foundations of international law (*Sparing Civilians*, OUP). Work on war received American Philosophical Association prize. ARC DECRA 2013. ASSA Panel D ECR award 2016. Lead CI on ARC DP on Ethics and Risk 2017. At front of debate on building probabilities into moral decision-making. Co-Editor of *Philosophers' Imprint*, a top philosophy journal. Chair of four PhD panels, co-authored paper in top journal with PhD student, mentored philosophy ECRs.

COLIN KLEIN (PHD+11)

Research featured in global media e.g. NY Times, Wired. Altmetric score in top 5%, multiple media appearances. Author of a prize-winning book on pain (MIT Press). Numerous publications in leading philosophy and psychology journals. ARC Future Fellow and ANU Futures Awardee. Founder of Australasian Society for Philosophy and Psychology. Has co-authored high-impact papers with PhD students and postdocs.

KATIE STEELE (PHD+11)

Has pioneered incorporation of moral reasons into decision theory. Forthcoming book on managing severe uncertainty. Substantial interdisciplinary and policy experience through work on climate change and economic theory. Associate editor of leading journal, *Philosophy of Science*. ANU Futures Awardee. Active collaborator with current and former PhD students and postdocs. Multiple co-authored papers and a prospective monograph with former PhD student.

LEXING XIE (PHD+13)

Leader of Computational Media Lab. Winner of Chris Wallace award for Outstanding Research 2017-18. Prominent work on ML algorithms in social media. Several patents. ARC, CRC, US Air Force, Department of Innovation grants totalling over $2m. Junior lab members have received multiple awards. Past members have gone on to roles in Twitter, Google, Baidu, as well as academic careers.

TONI ERSKINE (PHD+17)

Director of Coral Bell School. Award-winning book on duties to enemies and strangers in international politics and law. Associate Fellow of Centre for Future of Intelligence (Cambridge). Incoming Co-Editor of *International Theory*. Working with Google on AI. Prominent work on collective responsibility. Primary supervisor of 12 PhD students, many now tenured at top depts. Developed and ran mentoring programmes for ECRs in UK and Australia.

MARCUS HUTTER (PHD+22)

Developed the first and still only sound and complete theory of general AI. Numerous best paper awards, over $2m sole-CI basic research grants, recent USD250k grant from Future of Life institute on the control problem for general AI. First PhD student co-founded DeepMind. Subsequent postdocs and PhD students have gone on to prominent roles in DeepMind and other leading tech companies.

SYLVIE THIÉBAUX (PHD+23)

Co-editor in chief of *Artificial Intelligence*. Former lab director at NICTA. Multiple best paper awards. Project lead of Bruny Island Battery trial (multiple-award-winning project deploying AI). Influential research on optimisation and sequential decision theory for AI. Recent grants with Airbus, ARC, ARENA, US Airforce totalling over $4.5m. Has supervised 20 PhD students and 20 postdocs, who have gone on to academia, research labs (including Google, IBM, INRIA etc), or founded own startups.

BOB WILLIAMSON (PHD+28)

Fellow, Australian Academy of Science and Australian Mathematical Society. Lead author on *Technology and Australia's Future* (2015). Led a machine learning research group (2011-2015) reviewed as top five in the world. Co-authored NICTA bid, and as leader within NICTA contributed to raising apx $1b from government and industry. Has supervised 25 PhD students, and 50+ postdocs, who have gone on to elite universities and corporations (Amazon, Apple, Microsoft, Google, ElementAI, Facebook).

APPENDIX 2: GOVERNANCE STRUCTURE

HMI
HUMANISING
MACHINE
INTELLIGENCE