Yuhan Fu

E-mail: fu.philosophy@gmail.com

AoS: Moral Psychology, Philosophy of Cognitive Science, Philosophy of AI

AoC: Philosophy of Mind, Ethics, Film and Philosophy

Employment:

- **2024/04/11-Present Postdoctoral Research Fellow**
  Institute of Logic and Cognition, Department of Philosophy
  Sun Yat-sen University (Guangzhou, China)

Visiting Fellowship(s) and Affiliation(s):

- **2025/10/01- Present Visiting Fellow**
  Leverhulme Centre for the Future of Intelligence
  University of Cambridge
- **2023/07/01- Present Affiliated Researcher**
  Machine Learning and Normative Theory (MINT) Lab
  Australian National University

Education:

2018-2023 PhD in Philosophy University of Sheffield

- Supervisors: Ryan Byerly, Gerardo Viera, Stephen Laurence
- PhD thesis: Do we have a unified moral faculty? No (*Pass without corrections)*

2017-2018 MA in Film and Philosophy King's College London

2016-2017 MA in Philosophy King's College London

2009-2013 BA in Chinese Language and Literature South China Normal University

Peer-Reviewed Journal Articles:

1. **Fu, Y.,** Mei, Y. (Forthcoming): 'A Cautiously Optimistic View on Griefbot Technology'. *Phenomenology and the Cognitive Sciences.*
   - *Interdisciplinary research: Philosophy of AI, history of technology, concept acquisition, philosophy of grief, and animal thanatology*.
   - Drawing on rationalist account in cognitive science (Laurence & Margolis, 2024), animal thanatology, and the history of grief technologies, we argue that grief is underpinned by a domain-specific mechanism for processing irreversible relational loss in social relationships, and make a cautiously optimistic case that LLM-driven grief technologies (griefbots) function as alternative environmental inputs to grief performance (more individualised grieving expressions) rather than disruptions of the grief mechanism.

2. **Fu, Y.,** Mei, Y. (Forthcoming): 'Moral Flexibility without Mutual Benefits: From Change to

Disagreement', *Behavioral and Brain Sciences.*

- *Interdisciplinary research: moral psychology, social epistemology and science of belief.*
- We argue that moral flexibility often produces persistent disagreement rather than convergence, because social identity, confirmation bias, and emotion tend to block renegotiation and mutual agreement.

3. **Fu, Y.,** Viera, G. (2023): 'Puritanical norms do not stem from cooperative concern', *Behavioral and Brain Sciences.*
    - *Interdisciplinary research: moral psychology, cognitive science, and anthropology*.
    - We argue that puritanical norms cannot be explained solely by cooperation needs. Anthropological and archaeological evidence shows that self-indulgent behaviours (such as drinking, eating, and feasting) often enhance cooperation by reinforcing group identity.

Work-In-Progress Articles:

Interdisciplinary Research – Philosophy of AI, Cognitive Science and Moral Psychology

1. 'Beliefs about AI Systems are Identity-Constituting' (Under Review, *AI & Society*)
    - Single authored paper. Submitted to *AI & Society* in Jan 2026.
    - I argue that beliefs about AI operate across three interconnected dimensions: instrumental, psychological, and sociorelational. Therefore, unlike beliefs about other technologies, beliefs about AI are identity-constituting. They are guarded by mechanisms that protect self-identity and define in-group/out-group boundaries, making them resistant to revision and likely to crystallise into distinct ideological camps.
2. 'Superintelligent Robots and the Prudential Value of Digital Immortality' (Final Draft Finished)
    - Single authored paper. Under proofread.
    - To be Submitted to Special Issue "Superintelligent Robots" on *Philosophical Studies* by April 15th, 2026.
    - I target the transhumanist claim that future technologies, such as the emergence of Artificial Superintelligence (ASI), can radically improve human life by defeating diseases and aging. Focusing on digital immortality via ASI-assisted mind-uploading, I argue that digital immortality would not be prudentially good for individuals: the body is an irreplaceable source of well-being, and the relationship between enhancement, longevity, and well-being is non-linear.

Interdisciplinary Research – Philosophy, Moral Psychology and Cognitive Science

3. 'How Can I Make This Kind of Stupid Mistake' (First Draft Writing)
    - Single authored paper.
    - Target journal: *Analysis.*
    - This paper targets the phenomenon of slips: mistakes we make while acting with an explicit intention to achieve a goal. I challenge Amaya's (2013) claim that slips are intentional actions that violate the principle of revealed preference. Integrating psychological literature on error monitoring, I argue instead that slips are non-intentional by-products of a monitoring gap inherent in the architecture of skilled, habitual agency.

Moral Philosophy

4. 'Norms of Standing in Case of Praise' (First Draft Finished)
    - Single authored paper. Target journal: *The Philosophical Quarterly* (to be submitted by May 15th, 2026).
    - I argue that standing to praise and moral wrongfulness have at most a contingent relationship. Against recent accounts that extend the standing-to-blame framework to praise (Telech, 2024; Lippert-Rasmussen, 2022), I show that blame and praise are structurally asymmetric, and that a praiser's lack of standing does not thereby make their praise morally wrongful. When standingless praise seems problematic, this is better explained by insincerity, bad character, or relational context.

Ongoing Collaborative Projects:

1. Experience and AI "Character" Building (July 2025-):
    - Project description: This study investigates how diverse experiences shape machine personality and influence problem-solving. It employs continued pre-training to expose models to domain-specific texts in an unsupervised manner, simulating the accumulation of experience.
    - Project collaborator: this project is led by Dr Xi Wang (Speech and Language Processing) at the School of Computer Science, University of Sheffield
    - Role: Provide theoretical framework on the extent to which human personality is shaped by experience.

2. Large Language Models' moral characters (September 2025 -):
    - Project description: We examine moral preferences across seven different Large Language Models. We let each model role-play 40 different "persons", answer moral questionnaires, and analyse whether LLMs exhibit stable moral profiles, how these profiles vary across models, and whether role-played personas produce systematic shifts in moral judgement.
    - Collaborators: led by Socio-Cultural and Affective Neuroscience (SCAN) Lab at the Department of Psychology, Sun Yat-sen University
    - Role: Collect data and provide a philosophical perspective on moral preference in the human case.

3. How to train LMs to better understand human emotions in conversations (February 2026-):
    - Project description: Current Language Models are poor at detecting emotional cues in conversation. This project develops methods to improve LMs' sensitivity to affective signals in dialogue, with the aim of producing models that can more accurately recognise, interpret, and respond to emotional content in natural conversational contexts.
    - Collaborator: collaborate with Dr Mingjie Chen (speech processing, NLP), School of Computer Science, University of Sheffield.
    - Role: Conduct a comparative study on how humans detect affective cues in everyday conversation (this collaborative project has just started recently; more roles will be determined soon!)

Peer-reviewed Conference Presentations:

1. Annual Meeting of European Society for Philosophy and Psychology
   Moral Judgement and Psychological Experiments: A Case Study on Experiments with Psychopaths
   2019, Athens, Greece
2. Annual Meeting of Society for Philosophy and Psychology
   Poster: *The Logic of Universalisation Does Not Guide Moral Judgements*
   2021, Online
3. Metaethics in Society conference
   Can AI make Moral Judgements
   2022, the University of Nottingham, Nottingham, UK
4. PERITIA: Ethics of Trust and Expertise Conference
   Can We Trust AI's Moral Judgements
   2022, Yerevan, Armenia
5. Philosophy, AI and Society Doctoral Colloquium
   Whether AI systems can engage with moral cognition?
   2023, University of Oxford, Oxford, UK
6. Understanding Value XI Conference
   Puritanism, Cooperation and Moral cognition
   2023, the University of Sheffield, Sheffield, UK
7. The Ethics And Technology Early Career Group Workshop
   Algorithmic Biases, Discriminations, and Moral Responsibility
   2023, Online (Hosted by University of Vienna)
8. Formalising Responsibility: A Philosophy and Computer Science Workshop on notions of responsibility to support artificial agent decision making
   AI agency and moral responsibility
   2023, the University of Manchester, Manchester, UK
9. Ethics of Increasing AI Capabilities
   Technologies and the Grieving Process
   2024, Hannover, Germany
10. Annual Meeting of European Society for Philosophy and Psychology
    The Acquisition of Normative Concepts in Humans and Language Models
    2026, Warsaw, Poland

Invited Talks:

1. Heng-Seng Cognitive Science Talk Series
   AI, Moral Judgement and Moral Agency
   2022, Sheffield, UK
2. Explaining Normativity Workshop
   Troubles for Moral Sentimentalism
   2024, Sheffield, UK
3. Moral Realism and Its Epistemic Dimension
   Do we have a unified moral faculty?No
   2024, Guangzhou, China
4. Philosophical Conceptions of Autonomy in AI-Human Teaming Workshop

Artificial Superintelligence and Digital Immortality
2026, Sheffield, UK

Teaching and Mentoring Experience:

1. Graduate Teaching Assistant (2019-2023): delivered tutorial classes, organised 1-1 discussions about the course material, including in my office hours, and marked assessments for the following modules:
   - *'Mind, Brain and Personal Identity'*---- 2019-2020 Autumn; 2021-2022 Spring.

Module Topic(s): Philosophy of Mind, Philosophy of Psychology,

   - '*Philosophy of Science'* ----2019 Autumn.

Module Topic(s): Philosophy of Science, Epistemology.

   - *'Death'* ---- 2019-2020 Spring.

Module Topic(s): Metaphysics, Ethics.

   - *'Writing Philosophy'* ----2019-2020 Autumn.

Module Topics(s): How to write philosophical papers.

   - '*Elementary Logic'* ---- 2019-2020 Spring.

Module Topic(s): Formal Logic.

   - *'Reason and Argument'*----2019-2020 Autumn; 2021-2022 Autumn.

Module Topics(s): Logic, Arguments Constructions.

   - *'Matters of Life and Death'* ----2022-2023 Autumn.

Module Topic(s): Ethics, Political Philosophy.

2. Guest Lecturing for module *Philosophy and Society* (Oct, 2024)
   - Lecture 1: *Medical Health Welfare in China*
   - Lecture 2: *Is That Really an Involution? Comparison with In-group Member*

Module Topic(s): Social Philosophy, Moral Psychology, and Practical Ethics

3. Sheffield University Foundation Year Supervision:
   - Held Monthly Meeting, Provide Feedback, and Marked Assessments
4. Academic Mentor & Pastoral Support (Voluntary):
   - Provided dedicated guidance to international students (with a focus on Chinese cohorts) to help them navigate UK higher education.
   - Advised prospective and current students on structuring and refining postgraduate research proposals.
   - Delivered essential wellbeing support to help students manage academic pressure and cultural transitions.

Research Plan (Oct 2024 – Sep 2029):

Major Project (Interdisciplinary): *Matters of Life and Death in the Age of Artificial Intelligence*

- Project description: how current and future AI technologies will shape how we understand life, death and our well-being.
- Research outputs: 6 papers, and possibly one grant application.
- Timeline:
  - √October 2024 – September 2025:
    √ Literature review on philosophy of death, well-being, philosophy of AI
    √ Paper 1 on griefbots (submitted and accepted).

- October 2025 – September 2026
  √ Paper 2 on moralising beliefs about AI systems (submitted and under review)
  - Paper 3 on future technologies and immortality (final draft finished).
  - Paper 4 on psychology of grief (extending arguments from paper 1, paper outline done).
- October 2026 – September 2027
  - Paper 5 on the social role of AI. (depending on the development of AI technology): whether AI has its own social community, and whether they can be perceived as members in the human social and moral community.
- October 2027 – September 2028
  - Paper 6 on cross-cultural studies on how willing people are used to up-to-date technologies to navigate social life, especially in the context of grieving (Title TBD)
- October 2028 – September 2029
  - Explore book proposal drawing on the six papers.

Moral psychology and Moral Philosophy

- Research outputs: 4-5 papers.
- Timeline:
  - √October 2024 – September 2025
    √Paper 1 on moral disagreement (submitted and accepted)
  - October 2025 – September 2026
    - Paper 2 on slips and error monitoring (first draft writing)
    - Paper 3 on norms of standing in praise (final draft finished)
  - October 2026 – September 2027
    - Paper 4 on moral learning and whether we are moral Bayesians (extending PhD thesis, title TBD)
  - October 2027 – September 2028
    - Paper 5 on moral change (title TBD)
  - October 2028 – September 2029: TBD

Note: Timeline is subject to adjustment depending on teaching commitments.

Academic Workshop and Conference Organisation:

1. Co-organising Understanding Value 9 Conference at the University of Sheffield, 2020

Main roles: setting up online space for participants, writing applications for grants, managing conference email, collecting submissions, organising abstract reviewing, sending out acceptance and rejection emails, and chairing ~~some~~ parallel sessions.

2. Editor Manager for Centre for Engaged Philosophy at the University of Sheffield, 2021

Main roles: selecting papers for the CEP journal, managing social account and updating website.

3. Co-organising workshop: Imagining a Climate Crisis Curriculum at the University of Sheffield, 2022

Main roles: designing poster for the workshop, dinner booking, and bursary-relevant jobs.

Awards and Grants:

1. Chinese Students Awards by Great Britain China Educational Trust 2021 (£1500)
2. The Fundamental Research Funds for the Central Universities 2024, China (RMB 25,000, roughly £2800)

Research Impact beyond Academia

1. Participate in Impunity Watch, a human rights organisation based in The Hague that focuses on transitional justice and victim participation. Invited to give a talk on how technology could support victims' participation in justice processes and memorialisation. (the talk was postponed due to the clash with my recent talk at Sheffield in March 2026, new date TBD).
2. Gave a talk on "What is philosophy?" for year 4 students at Xuefu Primary School, Shenzhen, China, November, 2024

Additional Service to the Department / Profession:

3. 'Philosophy at the Showroom' Project. Here Sheffield philosophers introduce a film and lead an after-screen discussion on the selected film at the local independent cinema. I presented Everything Everywhere All at Once in December, 2022.
4. Diversifying reading lists for the Department of Philosophy, the University of Sheffield.
5. Member of the committee for 'PhD Qualifying Examination' in Ethics and Moral Psychology at the Department of Philosophy, Sun Yat-sen University. Read, evaluated and assessed two chapters and one report from seven PhD candidates who work on ethics and moral psychology. June 2025.

For Reference Letters:

1. Professor Ryan Byerly : t.r.byerly@sheffield.ac.uk
2. Dr. Luca Barlassina: l.barlassina@sheffield.ac.uk
3. Professor Hu Liu: liuhu2@mail.sysu.edu.cn