David Gray Grant

(Formerly David Michael Gray)

Department of Philosophy
Harvard University
25 Quincy Street
Cambridge, MA 02138, USA

dggrant@fas.harvard.edu
www.davidgraygrant.com

Employment

Harvard University

2020–present:	Lecturer on Philosophy
2018–present:	Embedded EthiCS Postdoctoral Fellow in Philosophy
2017–2018:	Ethics Pedagogy Fellow, the Edmond J. Safra Center for Ethics
2017–2018:	Embedded EthiCS Graduate Fellow in Philosophy
2014–2017:	Teaching Fellow, Harvard Kennedy School of Government

University of Southern California

Summer 2017:	Summer Research Fellow, Center for AI and Society

Education

Ph.D. in Philosophy, Massachusetts Institute of Technology, 2018

Dissertation: "Ethics for Artificial Agents"
Committee: Alex Byrne (chair), Tamar Schapiro, Miriam Schoenfield, Milind Tambe (Harvard University)

B.A. in Philosophy (with honors), University of North Carolina at Chapel Hill, 2005

Areas of Specialization

Applied Ethics (especially ethics of AI), Philosophy of Science (especially computer and data science), Social and Political Philosophy

Areas of Competence

Moral Philosophy, Philosophy of Mind, Metaphysics, History of Philosophy

Publications

"Embedded EthiCS: Integrating Ethics Broadly Across Computer Science Education" (with Barbara Grosz et al.), *Communications of the Association for Computing Machinery*, 2019

"Ethics and Artificial Intelligence in Public Health Social Work," in *Artificial Intelligence and Social Work*, ed. Milind Tambe and Eric Rice (Cambridge University Press), 2018

Dissertation Abstract

Machine ethics is an interdisciplinary field that focuses on ethical issues involved in the design of artificial agents–robots and other autonomous software agents. These systems have the capacity to make decisions about how to act in novel situations that were not specifically envisioned by their designers. The first part of my thesis argues against a prominent theory about how artificial agents ought to be designed: the "agential theory" of machine ethics, which says that artificial agents ought to be designed to behave only in ways that would be permissible for a human agent to behave. The second part sets out and analyzes a case study in machine ethics, focusing on the development of an artificial agent to assist with the planning of a public health intervention. I use the case study to show how (in this case and others) the task of determining which possible courses of action are morally acceptable can be divided between a system and its human users.

Works in Progress

"Ethics for Artificial Agents" (under review)

"Explanation and Machine Learning" (draft available soon)

"The Moral Demand for Explainable AI" (in preparation)

"Preventing Algorithmic Discrimination" (in preparation)

Fellowships and Awards

Certificate of Distinction in Teaching, Harvard University, Spring 2016

Presidential Fellowship, Massachusetts Institute of Technology, 2008–2009

Worth Award for Excellence in the Undergraduate Study of Philosophy, UNC Chapel Hill, 2005

Grants

Responsible Computer Science Challenge: Stage I, Harvard University, 2019-20

- $150,000. Collaborator on "Embedded EthiCS: Module Design and Teaching Materials for Core CS Courses" (PIs: Barbara J. Grosz, Radhika Nagpal)

Other Relevant Employment

Senior Research Fellow in Digital Ethics and Governance, Jain Family Institute, February 2018–present

- Conduct research, lead a weekly work-in-progress seminar for Institute Fellows in Digital Ethics and Governance, and consult on Institute projects

Philosophy Faculty, North Carolina Governor's School, Summer 2015 and 2016

- Prepared and taught three introductory philosophy classes a week to three different groups of fifteen rising high school seniors for the duration of the five-week summer program

Presentations

Keynote Address for the 18th Biennial Meeting of the International Society for Justice Research, Lisbon, Portugal, July 2020

Comments on Yvvonne MacPherson and Kathy Pham, "Ethics in Health Data Science," Radcliffe Institute Exploratory Seminar on the Ethics of Technology at Work and in Public Institutions, Harvard University, January 2020

Comments on Robert Long, "Fairness in Machine Learning: Against False Positive Rate Equality," Northeastern Information Ethics Roundtable, Northeastern University, April 2019

"Ethics for Artificial Agents," Northeastern University AI and Data Ethics Group, March 2019

"Fair Machine Learning and Moral Philosophy," Harvard Center for Research on Computation and Society Lunch Seminar, November 2018

"Algorithms in the Criminal Adjudicative Process" (panelist), Boston University Workshop on Algorithms, Ethics, and Accountability, March 2018

"Machine Ethics in Practice," Jain Family Institute, February 2018

"Preventing Algorithmic Discrimination," MIT Department of Philosophy Work in Progress Seminar, December 2017

"Ethical Considerations in the Design of Network-Based Prevention Programs," University of Southern California Center for Articial Intelligence in Society, July 2017

"Ethics and Artificial Intelligence: An Overview," University of Southern California Center for Articial Intelligence in Society, June 2017

"Skepticism About Personal Identity," Australian National University Philsoc Seminar, July 2012

"Living to See Another Day: Fission Cases and Personal Identity," Australasian Association of Philosophy Annual Conference, July 2012

Comments on Andrea Onofri, "Two Constraints On A Theory Of Concepts," Harvard-MIT Graduate Conference, March 2012

Teaching

As Primary Instructor

Intelligent Systems: Design and Ethical Challenges, Harvard University, Spring 2020

- Co-taught with Milind Tambe (Harvard Computer Science), this course teaches computer science students how to apply artificial intelligence to real-world social problems, and how to address ethical challenges that arise through appropriate design techniques

Embedded EthiCS Modules

Developed and taught Embedded EthiCS modules for the following Harvard computer science courses:

- Systems Security (James Mickens), Fall 2018 ("The Ethics of Hacking Back")
- Programming Languages (Stephen Chong), Spring 2018 ("Ethics in Software Verification and Validation")
- Introduction to Computer Science II (Stuart Shieber), Spring 2018 ("Morally Responsible Software Development")
- Design of Useful and Usable Interactive Systems (Krzysztof Gajos, Ofra Amir), Spring 2018 and Spring 2017 ("The Ethics of Inclusive Design")
- Fairness, Privacy, and Validity in Data Analysis (Cynthia Dwork), Fall 2017 ("Algorithmic Fairness and Equality of Opportunity")
- Networks (Yaron Singer), Fall 2017 and Spring 2017 ("Facebook, Fake News, and the Ethics of Censorship")

- Data Science II (Hanspeter Pfister, Mark Glickman, Verena Kaynig-Fittkau), Spring 2017 ("Ethics and the Data Science Process")
- Big Ideas in Computer Science (Henry Leitner), Spring 2017 ("Electronic Privacy")

Undergraduate Supervision

Senior thesis supervision at Harvard University:

- Aaron Kruk, in progress
- Sam Oh, in progress
- Michael Bervell, *Click Here to Accept: Electronic Informational Privacy on Social Media*, 2018-2019
- Aaron Fogelson, *The Legality of Obscenity in Virtual Reality*, 2018-2019

As Teaching Fellow

At the Harvard Kennedy School of Government:

- The Responsibilities of Public Action (Christopher Robichaud), Fall 2017, Fall 2016, Fall 2015, Spring 2014
- Ethics in Public Life (Christopher Robichaud), Fall 2016, Fall 2015, Fall 2014
- Economic Justice (Christopher Robichaud), Spring 2016, Fall 2014

At Harvard University:

- Philosophy of Psychology (Guven Guzeldere), Spring 2016
- Ancient Philosophy (Russell Jones), Spring 2015

At the Harvard Extension School:

- Utopia and Dystopia in Fiction and Philosophy (Christopher Robichaud), Fall 2016
- Contemporary Topics in Political Philosophy (Christopher Robichaud), Fall 2013

At the Massachusetts Institute of Technology:

- Minds and Machines (Alex Byrne), Fall 2011
- Justice (Lucas Stanczyk), Spring 2011
- Philosophical Issues in Brain Science (Alex Byrne and Pawan Sinha), Fall 2010
- Bioethics (Caspar Hare and David Jones), Spring 2010
- Classics of Western Philosophy (Rae Langton), Fall 2009

As Writing Advisor

At the Massachusetts Institute of Technology:

- Classics of Western Philosophy (Rae Langton), Fall 2012

As Grader

At the Massachusetts Institute of Technology:

- Theory of Knowledge (Roger White), Spring 2012

Graduate Coursework

* = *taken as an undergraduate*

Ethics, Epistemology, and Political Philosophy

- Normative Ethics (Caspar Hare)
- *Reasons (audit, Dorit Bar-On and Ram Neta)
- *Recent Developments in Political Philosophy (Geoff Brennan)
- *Relativism (Dorit Bar-On and Jesse Prinz)
- Topics in Theory of Knowledge (Roger White)

Metaphysics and Mind

- Causation, Explanation, Confirmation (Bradford Skow)
- Metametaphysics (Stephen Yablo)
- *Metaphysics of Modality (audit, William G. Lycan)
- Philosophy of Memory (Alex Byrne)
- *Philosophy of Mind (William G. Lycan)
- Possibility and Content (Agustín Rayo)

History

- *Aristotle (C.D.C. Reeve)
- Locke on Personal Identity (independent study, Alex Byrne)
- The Rationalists (Jeffrey McDonough)
- *Wittgenstein (Heather Gert)

Other

- *Modal Logic (Keith Simmons)
- Proseminar I: Frege to Carnap (Agustín Rayo and Rae Langton)
- Proseminar II: Quine to Kripke (Alex Byrne and Roger White)

Service

Participant, Harvard Integrating Ethics Across Computer Science Curricula Workshop, October 2018
Participant, FAT* Data Science Ethics Education Workshop, February 2018
Graduate Student Representative, MIT Department of Philosophy, 2011 and 2012
Graduate Student Recruitment Committee, MIT Department of Philosophy, 2010 and 2011

References

Alex Byrne (MIT): abyrne@mit.edu
Tamar Schapiro (MIT): tamschap@mit.edu
Miriam Schoenfield (MIT): miriams@mit.edu
Alison Simmons (Harvard): asimmons@fas.harvard.edu
Jeff Behrends (Harvard): jbehrends@fas.harvard.edu
Milind Tambe (Harvard Computer Science): milind_tambe@harvard.edu

Dissertation Summary

In March of 2018, one of Uber's autonomous vehicles struck and killed a pedestrian. Throughout that same year, Facebook's automated systems failed to detect a coordinated effort to persecute the Rohingya minority in Myanmar, contributing significantly to a major humanitarian crisis. In both cases, an autonomous software system failed to behave as expected, resulting in tragic consequences.

My dissertation focuses on theoretical and applied problems in machine ethics, a relatively new subfield of computer ethics that focuses on the ethical challenges involved in designing artificial agents–robots and other autonomous software agents. These systems are capable of deciding what to do in novel situations even when they have not been given explicit instructions about how to proceed by their human operators. Machine ethicists ask both how, from a moral point of view, artificial agents ought to make these decisions, and how, from an engineering point of view, they can be designed to make decisions that way.

In the first part of my dissertation, I consider what moral standards apply to the behavior of artificial agents. According to a prominent view that I call the agential theory of machine ethics, the answer is straightforward: the same moral standards that apply to the behavior of human agents. To find how we ought to design an artificial agent to act in a given situation, we need only ask how a human agent would be obligated to act in that situation. I distinguish two motivations for this view. The first is that artificial agents are moral agents, and so are subject to the same moral standards as human agents. The second is that the agential theory yields plausible results in concrete cases like those mentioned above.

I argue that both motivations fail. The first fails because artificial agents are not moral agents. Being a moral agent requires the capacity to form, and be guided by, your beliefs about what you ought to do. It is deeply implausible that contemporary artificial agents such as autonomous cars and robot vacuum cleaners possess this capacity. The second fails because the agential theory often makes incorrect predictions about how an artificial agent ought to be designed to act, as I demonstrate with a series of counterexamples. For example, because patients respond very differently to human therapists and "virtual therapists" (as recent empirical research demonstrates), different standards of conduct are appropriate for the two types of agents. I conclude that answering questions about how artificial agents ought to be designed will require developing new, domain-specific moral frameworks that take into account the morally significant differences between human and artificial agents.

Once we determine what moral standards should govern the behavior of an artificial agent, a new problem arises: how can we ensure that the agent's behavior will respect those standards? The second part of my dissertation explores this question by developing and analyzing a case study based on research I conducted at USC's Center for Artificial Intelligence in Society. The case study focuses on the development of an artificial agent to help plan public health interventions targeting homeless youth in Los Angeles. I argue that interventions planned by the agent must respect two moral demands that sometimes conflict: maximizing population-level benefits, and respecting each individual homeless youth's right not to be harmed. After outlining a framework to guide tradeoffs between these two demands (drawing on the public health ethics literature), I consider how the agent might be designed to balance them appropriately (drawing on the AI safety literature). In brief, the strategy I suggest is to use a combination of (1) simple rules the agent can readily apply to rule out obviously unacceptable intervention plans and (2) "safety conditions" that prompt human users for further input in more difficult cases. The result is a "division of epistemic labor" between artificial agents (which can apply simple rules very rapidly) and their human users (who are far more sensitive to moral nuances). This simple but effective strategy generalizes to a wide range of applied problems in machine ethics.