Legend:

× = To be written	*5 pages*
≈ = Needs jiggling (was written for the paper, is outdated)	*19 pages*
✓ = Complete (except for polishing, entry and exit, etc.)	*32 pages*
	Total page count: 60 pages

1. Abstract (1 page) ×

2. Introduction (5 pages) ≈

   (a) Problem and Motivation

      i. Artifical agents are becoming increasingly autonomous and are making increasingly consequential decisions. These are ethical decisions. Thus, we should use the centuries of existing literature on ethics to help us help computers make them.

   (b) My Idea/Contributions

      i. I present an implementation of automated ethical reasoning according to Kantian ethics that is faithful to philosophical literature. (include AI ethics diagram)

      ii. In Section 3a, I make a philosophical argument for why Kantian ethics, a kind of rule-based deontology, is the most natural of the three major ethical traditions (deontology, virtue ethics, consequentialism) to automate.

      iii. In Section 4a, I formalize a philosophically accepted version of the categorical imperative in DDL.

      iv. In Section 4b, I implement my formalization in Isabelle/HOL. My implementation includes axioms and definitions that can prove that appropriately represented sentences are permissible, obligatory, or prohibited. It can also return a list of facts used in the proof and, in some cases, an Isar-style human readable proof.

      v. In Sections 5a and 5b, I demonstrate my system's power and flexibility by using it to produce nuanced answers to two well-known Kantian ethical dilemmas. I show that, because my system draws on definitions of Kantian ethics presented in philosophical literature, it is able to perform sophisticated moral reasoning.

      vi. In Section 4c, I present a testing framework that can evaluate how faithful an implementation of automated Kantian ethics is to philosophical literature. My testing framework shows that my formalization is more philosophically accurate than prior work. This testing approach can be generalized to evaluate any implementation of automated Kantian ethics and to perform test-driven development for automated ethics.

      vii. In Section 6a, I demonstrate how my system can be used to generate novel(?) philosophical insights, thus showing that computational tools can help philosophers better study philosophy.

3. Methods (10 pages) ✓

   (a) Why Kantian Ethics (7 pages)

      i. Consequentialism

      ii. Virtue Ethics

      iii. Deontology/Kantian Ethics

   (b) Dyadic Deontic Logic (2 pages)

      i. Some interesting axioms and operators

   (c) Isabelle/HOL (1 page)
      i. What it is, what it lets you do

4. Implementing a novel formalization of the categorical imperative (12 pages)

   (a) The Formalization (6 pages) ✓
      i. Defining a Maxim (adopting O'Neill's definition)
      ii. Practical Contradiction Interpretation (summarizing Korsgaard's argument)
      iii. The Formalization Itself

   (b) Testing Framework (6 pages) ≈
      i. Comparision to other attempts (control group and Kroy's prior formalization)
      ii. Presenting the tests
         A. For each test, I will present both its philosophical justification and the code that runs the test.

5. Applications (12 pages) ✓

   (a) Lying vs Joking (6 pages)
      i. Philosophical explanation of the argument for why the FUL prohibits joking
      ii. My Approach (using thin common sense facts and assumptions)
      iii. Code running the example

   (b) Murderer Example (6 pages)
      i. Philosophical explanation of the dilemma of the murderer at your door
      ii. My Approach
      iii. Code running the example
      iv. Discussion of the need for and challenge of automating common sense

6. Discussion (20 pages)

   (a) Computational Ethics (8 pages) ≈
      i. Example of the philosophical insight (well-formed maxims, potential application to philosophy of doubt)
      ii. Arguing that computational tools can be valuable to philosophers

   (b) Limitations (2 pages) ×
      i. The Need for Common Sense
      ii. Formulating an Input Maxim
      iii. The AI ethics diagram and what else is needed to use the system in practice

   (c) Is Automated Kantian Ethics Even Possible? (4 pages)

   (d) Related Work (4 pages) ✓

   (e) Conclusion (2 pages) ×
      i. The idea of a computer doing ethical reasoning is scary, but insofar as people are going to keep building increasingly autonomous machines, it's better that they mimic ethical behavior.

ii. This is a proof-of-concept, but given computational progress and society's recognition of the need for AI ethics, we will see lots of progress and maybe someday this can actually be practical and usable.

7. References

8. Appendix

   (a) Full implementation of DDL

   (b) Running the tests on the control group

   (c) Running the tests on Kroy's formalization

   (d) Extra code from custom formalization and tests

   (e) What kind of computational ethics is a good idea?