Ethical By Designer - How to Grow Ethical Designers of Artificial Intelligence (Extended Abstract)*

Loïs Vanhée[1] , **Melania Borit**[2]
[1]Umeå University, Umeå, Sweden
[2]UiT The Arctic University of Norway, Tromsø, Norway
lois.vanhee@umu.se, melania.borit@uit.no

Abstract

Ethical concerns regarding Artificial Intelligence technology have fueled discussions around the ethics training received by its designers. Training designers for ethical behaviour, understood as habitual application of ethical principles in any situation, can make a significant difference in the practice of research, development, and application of AI systems. Building on interdisciplinary knowledge and practical experience from computer science, moral psychology, and pedagogy, we propose a functional way to provide this training.

1 Introduction

Against the backdrop of the challenges posed by ethical dilemmas of Artificial Intelligence (AI), spanning from bias in AI systems [Lee, 2018] to manipulation of human judgement [Henriksen, 2019], *virtue ethics*, understood as an approach to normative ethics that emphasizes moral character in contrast to approaches that emphasize duties and rules (deontology) or consequences of actions (consequentialism) [Carr, 2008; Hursthouse, 2017], becomes increasingly important in the debate around the impact AI will have on society. Virtue ethics gain attention as current tertiary education seems to fail in developing professional ethics and social responsibility skills [Chang *et al.*, 2020], codes of ethics are not drivers of ethical behaviour in moral exemplars in computing [Huff and Furchert, 2014], and developers' compliance with the principles set out in the various ethical guidelines is poor [McNamara *et al.*, 2018]. While moving away from preaching rules to focusing on cultivating the developers' character dispositions and moral attitude is a sensible advice [Harris, 2008], how to follow it is not straightforward, either for educators or for learners. We believe that an interdisciplinary approach integrating knowledge and experience from computer science, moral psychology and development, and pedagogy can provide a way for "broadening the scope of action, uncovering blind spots, promoting autonomy and freedom, and fostering self-responsibility" [Hagendorff, 2020] - that is, training for *ethical behaviour*, understood as habitual application of ethical principles in any situation [Treviño *et al.*, 2006] (e.g., routinely record anonymisation procedures in data mining activities involving personal data).

Thus, to those interested in becoming the kind of AI systems developers that the society needs and to those willing to contribute with training such developers, we propose using the GEDAI framework - Growing Ethical Designers of Artificial Intelligence. Implementing this framework in teaching and learning practices will mark a shift from achieving ethics for design(ers) (i.e., action-restriction through strict regulation of practice) [Dignum, 2019] to achieving *ethics by designers* - that is, empower AI systems developers to act self-responsibly in situations where morally relevant decisions have to be made. In the following, we describe shortly this framework and the way in which we envisage its use (for more details, see the full paper).

2 The GEDAI Framework Principles

The GEDAI framework builds on four core *principles*:

A) Ethical Behaviour (EB) is a central concept in virtue ethics and it is rooted in the social condition and the human psyche [Rest *et al.*, 1986]. As such, GEDAI proposes growing EB using advances in the domain of moral psychology and development. Within this domain, the Four Component Model of Moral Behavior [Narvaez and Rest, 1995] is the most studied and applied. This model introduces four dimensions along which individual moral ability and behavior can be grown: moral *sensitivity*, moral *judgement*, moral *motivation*, and implementation, also referred to as moral *action*.

B) Teaching ethical behaviour to AI systems developers can be operationalised using Intended Learning Outcomes (ILOs). Being statements about what a learner will achieve upon successful completion of an instructional unit (IU), ILOs are expressed from the learners' perspective and are measurable, achievable, and assessable. GEDAI advocates for defining specific ILOs for learning ethical behaviour.

C) Ethical behaviour is a transferable skill and, as such, it can be integrated with the practices of teaching AI. We believe that teaching ethical behaviour can and should be smoothly integrated in regular AI IUs, be they individual sessions, modules, courses, or programs, as integration as

[*] The original paper is referenced as: Vanhée, Loïs, and Melania Borit. "Ethical By Designer-How to Grow Ethical Designers of Artificial Intelligence." Journal of Artificial Intelligence Research 73 (2022): 619-631.

a strategy for developing transferable skills is proven to be more effective in higher education as it is more representative of the real-life application of skills in the workplace [Cottrell, 2001]. However, the GEDAI framework is still applicable in cases where other approaches are used.

D) Learners construct ethical behaviour, meaning, habits, and expertise through relevant learning activities, while teachers' task is to set up a learning environment that supports these learning activities. As such, GEDAI uses the *constructive alignment* educational principle [Biggs and Tang, 2011] to the design of IUs that integrate AI and ethical behaviour teaching.

3 The GEDAI Framework Description

The components of the GEDAI framework and the relationship between them are visualized in Figure 1.

The main elements of the teaching and learning process depicted by the GEDAI framework are phases, actions, and inputs/outputs. These *phases* are: 1) Operationalizing ILOs; 2) Planning of IU; 3) Implementing activities; 4) Refining IU. Whereas being presented here in a sequential order for facilitating understanding, these actions and phases are to be *undertaken in loops*, where actions/phases can overlap or a specific action/phase can trigger a revision of the previous one(s) (i.e., intra- and interphases loops).

In **Phase 1 (Operationalizing ILOs)**, the framework user has the task to specify ILOs suitable for the respective IU starting from higher level ILOs (actions a and b in Figure 1).

In the case of EB ILOs (action b), the specification of IU-specific EB ILOs from the general literature is not clear. Thus, here we propose using the *Integrative Ethical Education Model* [Narvaez and Lapsley, 2008; Narvaez and Bock, 2014], further operationalised and supplemented with examples of assessment and activities in the *Ethical Expertise Model* [Narvaez, 2009; Narvaez and Lies, 2009; Narvaez and Endicott, 2009; Narvaez and Bock, 2009]. These educational models build on the Four Component Model of Moral Behaviour, explained above in Principle A, and are based on evidence that such behaviour can be fostered by training *ethical expertise* [Huff, 2014; Narvaez, 2010], which is best gained through a novice-to-expert approach that moves through several stages of instruction while blending well-educated intuitions and good reasoning.

In **Phase 2 (Planning instructional unit)**, the framework user has the difficult task to integrate the two sets of operationalised ILOs and specify ILOs for the respective IU, and contextualize these ILOs (action c). GEDAI chooses to use an *integrative strategy* to teaching ethical behaviour as a transferable skill (i.e., skills are developed and taught explicitly within the core discipline with equal emphasis given to transferable skills and technical abilities), as opposed to embedding (i.e., no direct reference is made to developing transferable skills and the emphasis is on promoting the development of technical 'know-how') or bolting-in (i.e., skills are developed independently of the core discipline, enabling the explicit development of learners' transferable skills) [Chadha, 2006]. Several inputs play a role in action c (items III-VII in Figure 1), and the user has to be skilled in combining knowledge from various domains.

At the end of Phase 2, the framework user has to make a plan of learning activities that has to undergo a feasibility check (action d) during which it is assessed whether the planned activities are aligned with the ILOs and are compatible with administrative constraints and with other teaching activities, in order to avoid *non-productive* repetition and to cover blind spots.

In **Phase 3 (Implementing activities)**, the framework user carries on teaching as in the case of any other IU, with the mention that input V becomes relevant when the IU is an AI course in which ethics content is being integrated, in contrast with input IV that becomes relevant when the IU is a dedicated ethics course within the AI discipline. Considering that the output of action c is a set of innovative integrated AI and EB ILOs, the framework user has to be aware that innovative assessment tasks have to be formulated in action f, to assess the achievement of these ILOs.

In **Phase 4 (Refining instructional unit)**, the framework user performs the usual action of estimating ILOs acquisition by the learners (action g) through, for example, correlating grades with learners' feedback. Regardless of whether ethical skills acquisition was graded or not in Phase 3, in Phase 4, in addition to action g, we propose performing an estimation of the impact of ethical training (action h), which usually also involves collecting baseline data.

If the framework user has chosen to use the Ethical Expertise Model [Narvaez and Bock, 2014] in action b, to our knowledge, there is no assessment tool that holistically addresses the four components of moral behaviour as described in Phase 1 above, except a self-scoring instrument developed for dental education [Chambers, 2011] and a questionnaire designed for veterinary students [Verrinder and Phillips, 2014]. However, tools exist for assessing all of the four individual components. These tools can be general or profession-specific. The user of the GEDAI framework can explore how such profession-specific tools can be adapted for AI education.

4 The Strength of GEDAI

The *strength* of the GEDAI framework lies in the following. GEDAI focuses on teaching and learning ethical *behaviour*, which can be more straightforwardly embedded in daily life than other ethics-related skills. This approach seeks to grow *practical techno-ethical competent* learners, that is, technically-able persons with the habit of using ethical skills when producing concrete technical contents (e.g., growing the habit of relating "training data" to "assessing bias" rather than jumping to counting the layers of the neural network). At the same time, GEDAI makes explicit all the necessary steps to achieve this integration, thus being more concrete or operational than other available solutions. The GEDAI framework provides high adaptability by relying on meso-level ”containers” that can be filled in with specific content depending on the context (e.g., what are the hot topics of the moment). This feature ensures the longer-term effectiveness of the framework, as it can adapt to the moving landscape of ethical issues and deontological rules. That being said, pos-

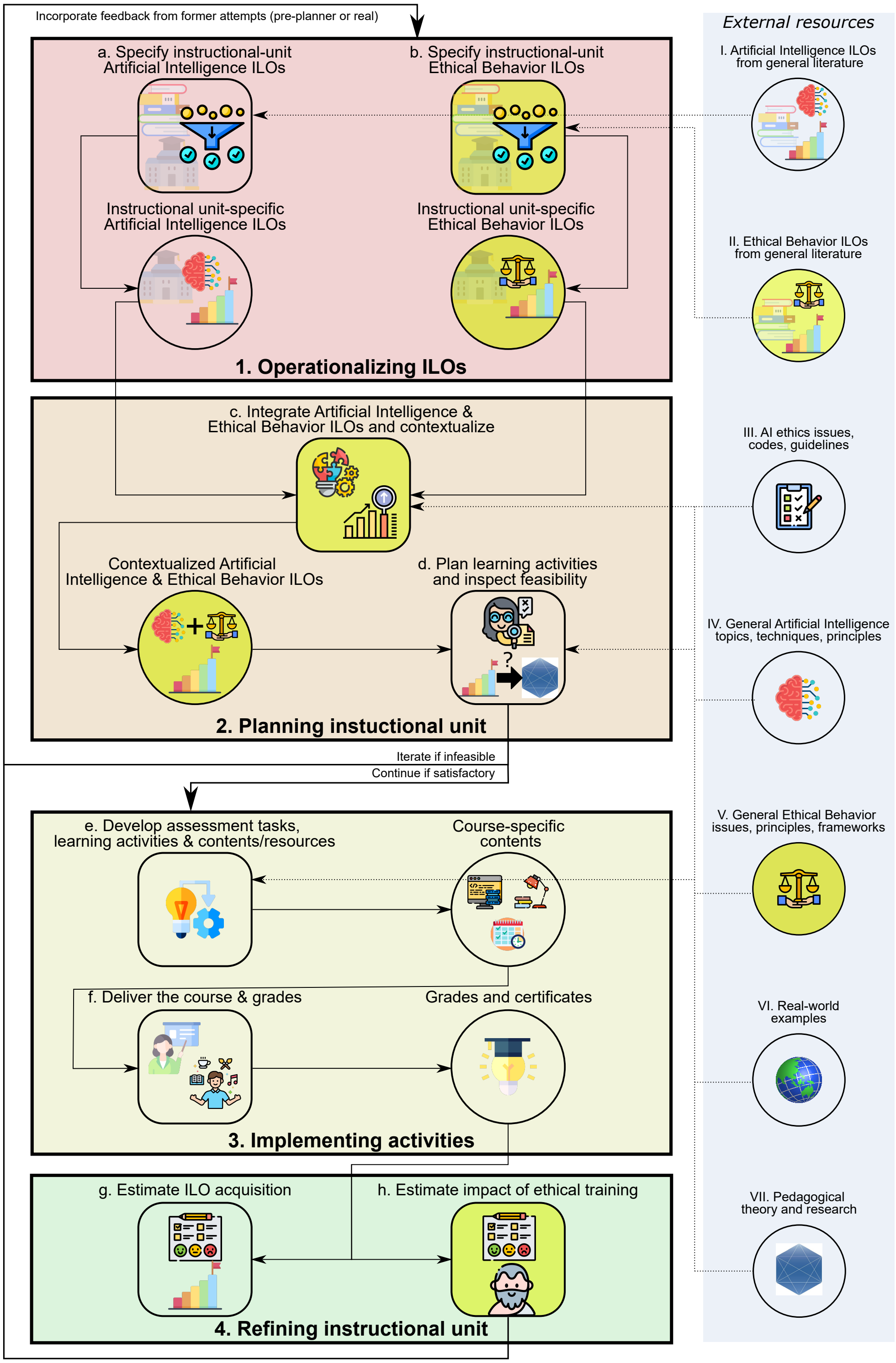
Incorporate feedback from former attempts (pre-planner or real)
a. Specify instructional-unit Artificial Intelligence ILOs
b. Specify instructional-unit Ethical Behavior ILOs
Instructional unit-specific Artificial Intelligence ILOs
Instructional unit-specific Ethical Behavior ILOs
1. Operationalizing ILOs
c. Integrate Artificial Intelligence & Ethical Behavior ILOs and contextualize
Contextualized Artificial Intelligence & Ethical Behavior ILOs
d. Plan learning activities and inspect feasibility
2. Planning instuctional unit
Iterate if infeasible
Continue if satisfactory
e. Develop assessment tasks, learning activities & contents/resources
Course-specific contents
f. Deliver the course & grades
Grades and certificates
3. Implementing activities
g. Estimate ILO acquisition
h. Estimate impact of ethical training
4. Refining instructional unit
External resources
I. Artificial Intelligence ILOs from general literature
II. Ethical Behavior ILOs from general literature
III. AI ethics issues, codes, guidelines
IV. General Artificial Intelligence topics, techniques, principles
V. General Ethical Behavior issues, principles, frameworks
VI. Real-world examples
VII. Pedagogical theory and research

Figure 1: GEDAI - Growing Ethical Designers of Artificial Intelligence framework. Squares represent actions; circles represent inputs/outputs. The yellow highlighted items represent the distinctive elements that integrate teaching ethical behaviour in the AI teaching and learning process. Intra- and interphases loops are not displayed.

sibly the strongest aspect of the GEDAI framework lies in its potential to facilitate the training of learners that have both technical know-how and the necessary ethical skills to *use* this know-how in the ”right” way; for example, when engaging with the task of defining the system’s objective function, the designer taught under the GEDAI framework will have the habit of remaining vigilant about direct, indirect, and unexpected involved stakeholders. However, as in the case of any other pedagogical tool or model, GEDAI only provides a foundation for organizing teaching and learning environments that maximize the chances for practical ethical expertise to be acquired, but cannot guarantee that this expertise will be demonstrated on the field.

As a final note, we have to mention that based on our own teaching experience and on numerous discussions with colleagues and learners, we are aware of the challenges posed by including yet another layer in the complex fabric of what has to be taught to a specific set of learners. Contrary to the general feeling of some educators, how to make space for this in our teaching without having to remove extensive AI disciplinary learning is a skill that we have to grow ourselves as teachers. However, this is not possible without the collaboration of the education leadership, the administration, the other teaching staff, and of the learners themselves. *The leadership* has to prioritize teachers’ growth time as educators over minimizing teaching costs, as quality student-oriented teaching adapted to the needs of our times requires more time and resources than the usual frontal teaching. *The administration* has to be able to adapt to the needs of implementing this teaching (e.g., adapted physical rooms, educational offers that adapt to the issues relevant in the society at a given time). *The other teaching staff* has to be open to have the same approach to their teaching as they have to their research. Thus, making use of learning analytics [Gašević *et al.*, 2017] and latest research in pedagogy should be the norm, not the exception. *The learners* have to see themselves as co-creators of value for themselves and for society from the moment they enter an educational program and not as dormant entities that will be activated after finishing a degree.

5 The Users of GEDAI

Learning professional ethics is acknowledged to benefit learners and professionals [Bebeau and Monson, 2008]. Thus, we envisage several user groups for the GEDAI framework: instructional units designers, industry, individual learners, researchers, and grant funders.

Designers of instructional units (teachers, course coordinators, program directors) who want to create a unit from scratch or update an existing unit can use the framework as a complementary tool to any other tools that are out there for instructional design. Moreover, they can use the framework to structure their critical reflection on choices to be made when integrating ethics in IU design and implementation. Since GEDAI is explicit about what external resources to feed into learning activities (items I-VII), it contributes to aligning activities in AI education with developments that happen in the real-world, thus aligning these activities with the needs of society and of learners. In terms of achieving the desired ethical behaviour, the framework helps with suggesting operational ILOs for this domain. Consistent application of the framework will help learners move on the continuum of moral behaviour expertise. As further aid, designers of IUs can build up on experiences of practitioners from other fields who have an integrative approach to training ethical behaviour.

Industry stakeholders interested in improving the capabilities of their workforce and in fulfilling their corporate social responsibility can use the framework in a similar way as the IU designers. Our framework provides a structure that can be customized for the individual needs of the company. Using the framework may reduce the transition costs towards ethical practices by allowing a smoother, incremental uptake into practice of the AI ethics guidelines by their employees.

Learners, both those learning by themselves and those enrolled in formal education, can use the GEDAI framework as an awareness-raising tool. By examining the elements of the framework, learners become aware of what is necessary to include in their self-growth process (e.g., various external resources), of the opportunity of learning ethics as a transferable skill together with learning the technical skills (much like learning statistics, for example), and of the need to estimate own ethical growth (action h) and measure their progression in becoming a responsible professional.

Funders of projects in education (e.g., the European Union) can use the framework for developing concrete funding programs and/or calls that promote the blending of ethics within operational skills, rather than as yet another requirement that ends up being presented as an appendix, while retaining the flexibility to incorporate what is relevant in the society at a certain point. The implementation of projects funded through such programs/calls would contribute to the growth of ethical behaviour skills in general and within the AI domain in a set-up that is more relevant to real-life and work-place situations.

Researchers interested in curriculum assessment can use the GEDAI framework for specifying content analysis codes suited for exploring how ethics are integrated in the current AI teaching practices or from a historical point of view. Those interested in the theoretical and methodological development of the field of teaching AI ethics can build upon the framework and possibly conceptualize/formalize it further. Researchers can use GEDAI for further framing the various contributions that can be made in teaching AI ethics (e.g., pairing activities and grading methods to ILOs).

6 Conclusions

Integrating abstract ethical recommendations and technical implementations is not a trivial task. Embedding ethical behaviour at the core of teaching and learning AI courses can help, and we propose drawing on expertise in computer science, moral psychology and development, and pedagogy to crack this hard nut. Being able to develop AI focused on social good now as well as in the future requires growing developers who behave ethically as a habit, even in the absence of an explicit set of rules, duties, or imperatives. Such a habit development in the GEDAI designers will have a long-ranging positive influence on the impact AI can have on society.

Acknowledgements

The first author acknowledges the support of the Knut and Alice Wallenberg Foundation (project number 570080103), of the project AutogrAIde "A Student-Driven Interdisciplinary Hackathon on Whether and How to Automate Grading & Assessment" (project number 570002260, The Rådet för Artificiell Intelligens, Umeå University, Sweden), and of the project GEDAI "Growing Ethical Designers of Artificial Intelligence", Umeå University, Sweden. Special thanks go to Bas Testerink, Frank Dignum, and Maarten Jensen for their comments on the layout of Figure 1 and for https://www.flaticon.com/ for some of the icons used in the image. The second author acknowledges the support of the project AFO-JIGG "Service design thinking to improve welfare and product quality in the Norwegian small-scale fishing fleet" (project number 302635, Research Council of Norway) and of the CRAFT Lab - Knowledge Integration and Blue Futures, UiT The Arctic University of Norway.

References

[Bebeau and Monson, 2008] Muriel J Bebeau and Verna E Monson. *Guided by theory, grounded in evidence: A way forward for professional ethics education*. 2008.

[Biggs and Tang, 2011] John B Biggs and C. Tang. *Teaching for quality learning at university*. McGraw Hill education (UK), 2011.

[Carr, 2008] David Carr. *Character education as the cultivation of virtue*. Routledge New York, NY, 2008.

[Chadha, 2006] Deesha Chadha. A curriculum model for transferable skills development. *Engineering Education*, 1(1):19–24, 2006.

[Chambers, 2011] David W Chambers. Developing a self-scoring comprehensive instrument to measure rest's four-component model of moral behavior: The moral skills inventory. *Journal of dental education*, 75(1):23–35, 2011.

[Chang *et al.*, 2020] Jen-Chia Chang, Hsiao-Fang Shih, and Kuang-Ling Chang. A Research on the Training Status of EECS Students' Core Competency in University of Science and Technology. In *International Conference on Industrial Engineering and Engineering Management*, pages 1069–1072. IEEE, 2020.

[Cottrell, 2001] Stella Cottrell. *Teaching study skills and supporting learning*. Palgrave Basingstoke, 2001.

[Dignum, 2019] Virginia Dignum. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer Nature, 2019.

[Gašević *et al.*, 2017] Dragan Gašević, Vitomir Kovanović, and Srećko Joksimović. Piecing the learning analytics puzzle: A consolidated model of a field of research and practice. *Learning: Research and Practice*, 3(1):63–78, 2017.

[Hagendorff, 2020] Thilo Hagendorff. The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, 30(1):99–120, 2020.

[Harris, 2008] Charles E Harris. The good engineer: Giving virtue its due in engineering ethics. *Science and Engineering Ethics*, 14(2):153–164, 2008.

[Henriksen, 2019] Ellen Emilie Henriksen. Big data, microtargeting, and governmentality in cyber-times. The case of the Facebook-Cambridge Analytica data scandal. Master's thesis, 2019.

[Huff and Furchert, 2014] Chuck Huff and Almut Furchert. Toward a pedagogy of ethical practice. *Communications of the ACM*, 57(7):25–27, 2014.

[Huff, 2014] Chuck Huff. From meaning well to doing well: Ethical expertise in the GIS domain. *Journal of Geography in Higher Education*, 38(4):455–470, 2014.

[Hursthouse, 2017] R Hursthouse. Virtue Ethics in Stanford Encyclopedia of Philosophy, Metaphysics Research Lab, Palo Alto, CA, 2017.

[Lee, 2018] D Lee. Amazon scrapped 'sexist AI'tool, BBC News, October 10, 2018.

[McNamara *et al.*, 2018] Andrew McNamara, Justin Smith, and Emerson Murphy-Hill. Does ACM's code of ethics change ethical decision making in software development? In *Proceedings of the 2018 26th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering*, pages 729–733, 2018.

[Narvaez and Bock, 2009] Darcia Narvaez and Tonia Bock. *EthEx: Nurturing character in the classroom. Book Two: Ethical Judgment*. Chapel Hill, NC: Character Development Publishing, 2009.

[Narvaez and Bock, 2014] Darcia Narvaez and Tonia Bock. *Developing ethical expertise and moral personalities*. Routledge New York, NY, 2014.

[Narvaez and Endicott, 2009] Darcia Narvaez and Leilani Endicott. *EthEx: Nurturing character in the classroom. Book One: Ethical Sensitivity*. Chapel Hill, NC: Character Development Publishing, 2009.

[Narvaez and Lapsley, 2008] Darcia Narvaez and Daniel Lapsley. Teaching moral character: Two alternatives for teacher education. *The Teacher Educator*, 43(2):156–172, 2008.

[Narvaez and Lies, 2009] Darcia Narvaez and James Lies. *EthEx: Nurturing character in the classroom. Book Three: Ethical Motivation*. Chapel Hill, NC: Character Development Publishing, 2009.

[Narvaez and Rest, 1995] Darcia Narvaez and James Rest. The four components of acting morally. *Moral behavior and moral development: An introduction*, 1(1):385–400, 1995.

[Narvaez, 2009] Darcia Narvaez. *EthEx: Nurturing character in the classroom. Book Four: Ethical Action*. Chapel Hill, NC: Character Development Publishing, 2009.

[Narvaez, 2010] Darcia Narvaez. Moral complexity: The fatal attraction of truthiness and the importance of mature moral functioning. *Perspectives on Psychological Science*, 5(2):163–181, 2010.

[Rest *et al.*, 1986] James R Rest, Muriel Bebeau, and Joseph Volker. *An overview of the psychology of morality*. Praeger New York, NY, 1986.

[Treviño *et al.*, 2006] Linda K Treviño, Gary R Weaver, and Scott J Reynolds. Behavioral ethics in organizations: A review. *Journal of management*, 32(6):951–990, 2006.

[Verrinder and Phillips, 2014] Joy M Verrinder and Clive JC Phillips. Identifying veterinary students' capacity for moral behavior concerning animal ethics issues. *Journal of veterinary medical education*, 41(4):358–370, 2014.