

Extending the Overarching Properties with Ethical Principles

Taylor A. Thoman

OSTEM Internship Program, NASA's Langley Research Center

This paper proposes an extension of the Overarching Properties (OPs) theory to include ethical principles for consideration during the lifecycle of a software system. This extended theory, the Overarching Properties and Principles (OP²s¹), is a framework for generating assurance artifacts for complex systems with potential operational safety and ethical consequences. The goal of this paper is explore the value in making historically implied ethical considerations explicit to assure stakeholders that their engineered systems are suitable for public deployment.

I. Introduction

The Overarching Properties (OPs) theory emerged from an effort to “streamline assurance processes” [1]. Assurance refers to the notion that the system has been designed and developed in such a way as to warrant some confidence — *assurance* — that the system will be suitable — *safe, secure, reliable, etc. enough* — to operate. The OPs theory distills the theoretical notion of assurance into three properties. This document presents an informal ethical analysis of the OPs theory and its properties to introduce and preliminarily address the question: as a complete representation of assurance, does the sufficient set of OPs ensure adequate coverage of *ethics*²? In other words, does demonstrating the possession of the three OPs in a system assure stakeholders that the product will be *ethical enough* for public deployment?

Part of the motivation for generalizing assurance processes is to advance engineering and approval methods for new, increasingly complex algorithmic technologies [3]. The novelty of these technologies introduces questions about their safety and ethical implications [4, 5]. Safety and ethics are nebulous concepts, particularly when applied in the realm of technology. This paper generally grounds the conceptualization of safety and ethics as relating to two domains: (1) process considerations during technology design and development and (2) assessed consequences of technology deployment and operation. This grounding means we can think of safety and ethics as either safe/unsafe or right/wrong ways of engineering systems, or as safe/unsafe or right/wrong deployments of technologies³.

In current practice, the safety and quality engineering disciplines do a good job of defining safe/unsafe processes and consequences [6]. These disciplines explicitly and systematically identify and assess the potential safety harms of deploying a system. The ethical side, however, tends to be only implied, not defined, in engineering and left to the discretion of the product developer and approver. There is no ethics engineering discipline that defines right/wrong engineering processes and consequences to explicitly evaluate potential ethical harms, at least not yet⁴.

This paper uses a thought experiment to explore if it is possible to make these ethical implications explicit and if doing so adds value to the assurance process. These topics are explored under an informal ethical analysis which proceeds as follows. Section II provides context on the established OPs theory. This section provides definitions directly from the OPs literature. Section III comments on the ethical coverage of the OPs. Here, the analysis reveals where ethical considerations are implied within the OPs. Section IV proposes that ethical considerations should sometimes be made explicit within the OPs theory. This section derives ethical principles to follow during the product lifecycle. Section V gathers the findings of the informal ethical analysis to comment on the value that ethical considerations may bring to the assurance process. Finally, the paper concludes by proposing the combined set of Overarching Properties and Principles (OP²s) as a framework for systematically generating artifacts to demonstrate both safety and ethical assurance.

¹Pronounced “Oh-Pee-Twos.”

²The most appropriate notion of ethics here seems to be Normative Ethics, the “branch of moral philosophy, or ethics, concerned with criteria of what is morally right and wrong. It includes the formulation of moral rules that have direct implications for what human actions, institutions, and ways of life should be like” [2]. Whether this branch of ethics is the appropriate usage remains to be proven. This paper only serves to make an initial connection between ethics and the assurance process. Deeper exploration into compatible notions of ethics and assurance is a next valuable step for further research.

³Later uses of the terms ‘safety’ and ‘ethics’ may be simplified to appear standalone but still generally refer to one or both of these meanings.

⁴Engineers have not been ignoring ethics; the teaching and application of ethics has just not been systematized in the same way most everything else has been.

II. Background

This paper builds upon the growing body of literature that introduces and explains the Overarching Properties (OPs) theory and its nuances in depth [1, 7–11]. The OPs theory describes an idea for a sufficient set of properties that, if demonstrated as possessed by a system, could provide enough grounds for confidence that the system is justifiably safe for approval and deployment. The main source of truth for the OPs, *Understanding the Overarching Properties* [1], proposes three labeled properties — **Intent**, **Correctness**, **Innocuity** — to represent this sufficient set. The property labels provide only a convenient shorthand; all of the semantic content is in the text. The full description from *Understanding the Overarching Properties* is shown below. The rest of this paper uses these terms according to these definitions.

Properties

- a. **Intent**: The *DEFINED INTENDED BEHAVIOR* is correct and complete with respect to the *DESIRED BEHAVIOR*.
- b. **Correctness**: The *IMPLEMENTATION* is correct with respect to its *DEFINED INTENDED BEHAVIOR*, under *FORESEEABLE OPERATING CONDITIONS*.
- c. **Innocuity**: Any part of the *IMPLEMENTATION* that is not required by the *DEFINED INTENDED BEHAVIOR* has no *UNACCEPTABLE IMPACT*.

Definitions

- a. *DESIRED BEHAVIOR*: Needs and constraints expressed by the stakeholders (this includes those needs and constraints identified by the *SAFETY ASSESSMENT* and those mandated by regulations).
- b. *DEFINED INTENDED BEHAVIOR*: The record of the *DESIRED BEHAVIOR*.
- c. *IMPLEMENTATION*: *ITEM* or combination of inter-related *ITEMS* for which acceptance or approval is being sought.
- d. *ITEM*: a hardware or software element having bounded and well-defined interfaces.
- e. *FORESEEABLE OPERATING CONDITIONS*: External and internal conditions in which the system is used, encompassing all known normal and abnormal conditions.
- f. *UNACCEPTABLE IMPACT*: An impact that compromises the *SAFETY ASSESSMENT*.
- g. *SAFETY ASSESSMENT*: The systematic identification of *FAILURE CONDITIONS* and classifications in an operational context, evaluation of the architecture against safety objectives arising from these hazards, evaluation of potential common modes and threats, defining additional intended behaviors to support claims within these evaluations and showing that the safety objectives are satisfied by the *IMPLEMENTATION*.
- h. *FAILURE CONDITION*: “A condition having an effect on the aircraft and/or its occupants, either direct or consequential, which is caused or contributed to by one or more failures or errors, considering flight phase and relevant adverse operational or environmental conditions or external events.”

Requisites *for showing possession of the Overarching Properties*

- a. *DEFINED INTENDED BEHAVIOR* exists.
- b. *FAILURE CONDITIONS* are defined.
- c. The record of the *SAFETY ASSESSMENT* exists.
- d. The record of the *FORESEEABLE OPERATING CONDITIONS* exists.
- e. The *IMPLEMENTATION* exists.
- f. Development Assurance Level (DAL) assignments based on *FAILURE CONDITION* classifications exist.

Assumptions *which need only to be stated, not justified*

- a. Stakeholders have the knowledge to express the *DESIRED BEHAVIOR*.
- b. Performing *SAFETY ASSESSMENT* is not covered by these Overarching Properties.

Constraints *on how Overarching Property possession must be demonstrated*

- a. The process to ensure possession of the Overarching Properties must be defined and conducted as defined.
- b. The means by which the *DEFINED INTENDED BEHAVIOR* is shown to be correct and complete is commensurate with the DAL.
- c. Criteria for evaluating the artifacts are defined and shown to be satisfied individually and collectively.
- d. All artifacts are under configuration management and change control.
- e. When tiers of decomposition are used, the means of showing correctness among the tiers and to the *DEFINED INTENDED BEHAVIOR* must be defined and conducted as defined.
- f. The *IMPLEMENTATION* must be correct when functioning as part of the integrated system or in environment(s) representative of the integrated system.

- g. All design and manufacturing data to support consistent replication of the type design and instructions for continued airworthiness must be established.
- h. The *SAFETY ASSESSMENT* must address all of the *IMPLEMENTATION*.

III. Overarching Properties

Ethics are implied in most things. Despite whether people explicitly document them, ethical considerations exist in engineering. An engineered system itself may not be intrinsically ethical or unethical⁵, but the way the system is or was conceptualized, designed, developed, implemented, deployed, and/or operated could be ethical or unethical. This concept is like the notion of safety in technology. An algorithm alone, un-executed, is neither safe nor unsafe. However, take the following example.

Example: Deployment of an algorithm that takes control away from the pilot and causes the aircraft to crash.

The deployment of an algorithm that takes control away from the pilot and causes the aircraft to crash is an *unsafe deployment* of the algorithm.

Thus, considering ethics in engineering is conceptually similar to considering safety in engineering. This similarity may explain why safety is often used as a proxy for ethics in engineering. There may exist a notion that considering safety in engineering is sufficient for considering ethics. If the product is safe enough, then it is also ethical enough. While similar, safety and ethics do not map directly onto the same concerns. The safety analysis is concerned with the fact that the algorithm may cause the aircraft to crash. The ethical analysis is concerned with the fact that the algorithm may take control away from the human pilot⁶.

Example: Deployment of an algorithm that takes control away from the pilot and causes the aircraft to crash.

ethical concern

safety concern

Both types of concerns are relevant and should be distinguished and mitigated to an acceptable level when designing advanced algorithmic systems⁷. So far in engineering, the safety consequence has provided a locus of control upon which to direct the assessment of suitability. What might we glean by also directing resources to focus on the ethical consequence, like a control authority dispute or gap, not just as a contributing factor to a safety consequence, but as a locus of control for assessment itself?

Looking at the ethical concern in the example, we may observe that deploying an algorithm that takes control away from the pilot and causes the aircraft to crash is not only an *unsafe deployment* of the algorithm but also an *unethical deployment* of the algorithm. Considering safety as a proxy for ethics would only work if we can be sure of at least two things. First, the safety assessment process is robust to uncovering all the potential causes of the ethical concerns which pose the risk of unethical deployment, in addition to the safety concerns which pose the risk of unsafe deployment. Second, we can guarantee that all ethical concerns will map to safety consequences, without any non-safety related ethical consequences.

More likely, there may be an inherent notion in engineering that ethics are implied as covered by the good intentions of the product developers and societal norms which are reinforced by regulations. An optimistic position is to assume all engineers want to build ethical, safe systems and all approvers only want to approve ethical, safe systems. Still, just as the safety engineering discipline emerged and evolved⁸ from implied safety in ‘good engineering’ to defined safety engineering best practices, so too may the ethics engineering discipline need to emerge and evolve⁹. Making safety considerations explicit in engineering made our systems better without invalidating the optimists viewpoint. Perhaps the same can happen for ethics.

To test these intuitions, let us begin a thought experiment considering ethics in a similar manner to the way that we consider safety in engineering¹⁰. The thought experiment begins with turning these assumed notions above into hypotheses.

⁵Some philosophers may disagree with this position. That debate is out of scope for this paper but may be an interesting consideration in future research.

⁶Two very simplified, non-exhaustive statements about the concerns, purposes, motivations, etc. of these analyses.

⁷These types of concerns, ‘ethical concern’ and ‘safety concern,’ are used conceptually as buckets to distinguish between principles, mechanisms, and harms of a safety nature and an ethical nature, rather than the principles, mechanisms, and harms themselves. This notion is preliminarily and abstractly addressed in Section IV.B. Future research may explore how to properly assess each type of item within the ethical or safety concern.

⁸As iterative reactions to near-misses, mishaps, and/or tragic accidents.

⁹Hopefully as a proactive call-to-action in the field.

¹⁰A simple solution may be to add ethical harms to the safety assessment. The point of this exercise is to encourage readers to consider whether that simple solution is sufficient or if ethical harms should be examined separately.

Ethics are not explicitly considered as a discipline in engineering because...

Hypothesis A: Considering safety in engineering is sufficient for considering ethics.

Hypothesis B: Ethics are covered by the good intentions of the product developers and societal norms reinforced by regulations.

We can proceed to evaluate the OPs for ethical coverage with these hypotheses in mind. Later, we will gather our findings and consider their implications with respect to the hypotheses.

A. Ethical Coverage of the OPs

Upon examination of the stated intention of the OPs theory, it is clear that ethics are implied. In the philosophy section of *Understanding the Overarching Properties*, Holloway writes, “You can know for certain that the product is *suitable*. Thus, if you are charged with deciding to approve or disapprove the product, you can confidently approve it, without any fear of making the wrong decision” [1]. Using phrases like ‘suitable product,’ ‘confident approval,’ and ‘without fear of making the wrong decision’ evokes the underlying moral¹¹ and ethical assurances inherent in the intentions of the OPs theory.

The properties inherit this implied ethical coverage and provide the framework for generating assurance artifacts which in turn include implied ethical elements. What makes the idea of a product suitable for approval is the implied level of safety and ethics in the **Intent** of the product, the first property of the OPs. The statement of **Intent** reads as,

The *DEFINED INTENDED BEHAVIOR* is correct and complete with respect to the *DESIRED BEHAVIOR*.

The implications here are that the *DESIRED BEHAVIOR* is ethical and safe and the *DEFINED INTENDED BEHAVIOR* properly represents the ethical, safe *DESIRED BEHAVIOR*. In fact, the term *DESIRED BEHAVIOR* is defined as,

Needs and constraints expressed by the stakeholders (this includes those needs and constraints identified by the *SAFETY ASSESSMENT* and those mandated by regulation).

The word ‘needs’ is used loosely to include both needs and wants. Reading between the lines, we may be able to reasonably claim that the nested notions inside the **Intent** statement support both **Hypothesis B**, that ethics are implied in engineering as natural to the wants and needs of the stakeholders and regulations, and **Hypothesis A**, that ethics are otherwise considered sufficiently covered by safety considerations.

Despite being the most technical of the properties, the next property, **Correctness**, also includes embedded ethical considerations. The statement of **Correctness** reads as,

The *IMPLEMENTATION* is correct with respect to its *DEFINED INTENDED BEHAVIOR*, under *FORESEEABLE OPERATING CONDITIONS*.

This property is meant to be straightforward: the product is a correct manifestation of the requirements. In engineering, sometimes the notion of being straightforward is conflated with the term ‘objective,’ and sometimes that means the term ‘subjective’ is considered undesirable. This negative connotation of subjectivity may be why some engineers are uncomfortable with recognizing and naming non-objective considerations in engineering processes¹². Still, ethics are implied in the notion of **Correctness**, in the inherited ethical assumptions within the *DEFINED INTENDED BEHAVIOR* and also in the determination of *FORESEEABLE OPERATING CONDITIONS*. We already explored the ethics embedded in the notion of intended behavior and Holloway directly addresses the epistemic¹³ challenges of claiming completeness of *FORESEEABLE OPERATING CONDITIONS* [1]. Ethical considerations influence this determination, just as importantly as safety considerations do. We will explore these ideas in the next section.

The final property, **Innocuity**, clearly evokes ethical implications with its statement:

Any part of the *IMPLEMENTATION* that is not required by the *DEFINED INTENDED BEHAVIOR* has no *UNACCEPTABLE IMPACT*.

¹¹ Another note for the philosophers: the terms ‘morals’ and ‘ethics’ may be used somewhat interchangeably and without exceptional semantic discretion throughout this paper. Generally, ‘morals’ is used where personal belief or behavior is implied and ‘ethics’ is used otherwise to refer to general rules of right/wrong. Further nuanced research may come to more accurately sharpen the distinction.

¹² Philosophers have been making these observations about objectivity in science for a long time [12, 13].

¹³ Meaning “Of or relating to knowledge, or to its extent, linguistic expression, or degree of validation” [14].

Two key definitions are nested within this property.

UNACCEPTABLE IMPACT: An impact that compromises the **SAFETY ASSESSMENT**.

SAFETY ASSESSMENT: The systematic identification of **FAILURE CONDITIONS** and classifications in an operational context, evaluation of the architecture against safety objectives arising from these hazards, evaluation of potential common modes and threats, defining additional intended behaviors to support claims within these evaluations and showing that the safety objectives are satisfied by the **IMPLEMENTATION**.

Holloway supplies brevity to the intent of this property, stating, “Nothing extra in the **IMPLEMENTATION** can negatively affect safety,” and, “... the meaning of [**Innocuity**] is ‘do no wrong things’ (where wrong means *unsafe*)” [1]. Again we may be able to take these statements and notions in support of **Hypothesis A**, that all unacceptable ethical impacts are sufficiently covered by those with safety consequences. Regardless of specification of safety language in the definitions, the spirit of this property provides implied ethical coverage. Ethical concerns with consequential safety impacts are sufficiently covered. The definition of **UNACCEPTABLE IMPACTS** can easily be expanded to include non-strictly safety-related ethical concerns, the same way that environmental, cybersecurity, or financial impacts are considered in engineering suitable systems.

IV. Overarching Properties and Principles

We have seen that it is possible to make explicit the ethical implications within the OPs theory as-is. Now, let us explore the idea that, in some cases, systematically making these ethical considerations explicit can add value to the assurance process. For some time now, the designs of algorithmic systems have become increasingly complex. Additionally, the operational environments in which these algorithmic systems are deployed are also becoming increasingly complex. Overall, system complexity becomes compounded because the algorithmic systems are often interconnected with and interdependent on the development and deployment environments. These complex systems-of-systems include varying levels of algorithmic autonomy, which is often the reason for requiring assurances that the systems are suitable for approval. The OPs provide a framework for generating safety assurance artifacts for these complex systems-of-systems deployments. Would it also be valuable to generate ethical assurance artifacts for these systems?

Remember our unsafe, unethical algorithm deployment example.

Example: Deployment of an algorithm that takes control away from the pilot and causes the aircraft to crash.

ethical concern

safety concern

In today’s complex systems-of-systems world, the safe deployment of that same algorithm does not depend solely on the development of the code for the algorithm. The safe deployment also depends on the technical publications informing the pilots of the functionality and control authority of the algorithm, the training the pilots receive to handle failure modes associated with that algorithmic system, even on philosophical considerations of whether or not the algorithm should be used in the flight control system, the organizational pressures and cultural issues present when designing the aircraft and the flight control system, and innumerable more factors.

These considerations all have consequential effects on safety, but very quickly overload or are considered out of scope of the system safety assessment. The central items of focus in a traditional suite of system safety analyses are failures, both functional and hardware, and software errors which cause hazards of potential risk. These hazards are typically analyzed to understand one step before, the cause, and one step after, the consequence. In a system-of-systems, there may be causal factors many steps before and consequences many steps after. Thus, things like organizational pressures and training gaps do not always end up in the safety assessment because they exist in the periphery of the hazard.

Current engineering practice attempts to address these overflow considerations using quality or human factors analyses or by thinking of safety as a systems-theory problem and implementing a safety management system for the organization. These practices add supplemental value to the assurance process. However, they may not map directly onto the same ethical concerns, potentially violating **Hypothesis A**. As part of the thought experiment, let us consider any overflow considerations that are not covered by safety, quality, or human factors engineering processes to be ethical considerations. Systematically considering these elements may in turn add similar supplemental value to the assurance process.

Using this ethical lens, we are now in need of a method for systematically addressing the ethical concerns illustrated by the unsafe, unethical algorithm deployment example. The OPs theory is a tool which provides an explicit method

of addressing the safety concerns and assuring the safety of the system. We can continue the thought experiment by leveraging the OPs theory to provide explicit coverage of ethical assurance, in addition to safety assurance. One way to do this exercise is by extending the OPs to include ethical principles as a supplemental extended theory called the Overarching Properties and Principles (OP²s).

A. Properties and Principles

An important distinction must be made regarding the language used in the OP²s proposal. The Overarching *Properties* are aptly named because they are measurable¹⁴ characteristic qualities within the system. **Intent** is measurable via correctness and completeness analysis of the *DEFINED INTENDED BEHAVIOR* with respect to the *DESIRED BEHAVIOR*. **Correctness** is measurable via correctness and completeness analysis of the *IMPLEMENTATION* with respect to the *DEFINED INTENDED BEHAVIOR*. **Innocuity** is measurable via correctness and completeness analysis of the results of the *SAFETY ASSESSMENT* with respect to the *FORESEEABLE OPERATING CONDITIONS*. Safety is measurable in a system via evaluation of the results of the *SAFETY ASSESSMENT* against industry-accepted definitions of risk tolerance thresholds and acceptance levels.

These tolerance thresholds and acceptance levels may not be appropriately analogous for ethical considerations, which is another distinguishing factor between safety and ethics. Instead of proposing ethical characteristic qualities (properties) to measure within a system, a more intuitive approach is to propose ethical principles which can be followed during the lifecycle of a system. Thus, the extended OPs theory is proposed as the Overarching Properties and *Principles* (OP²s), where the Overarching Properties include implicit ethical considerations and the Overarching Principles include explicit ethical considerations to bolster and reinforce the properties. The OP²s can then be satisfied by demonstrating the *possession* of the Overarching Properties and the *adherence* to the Overarching Principles.

B. Adding Ethical Principles to Overarching Properties

The main point of contention against accepting **Hypothesis A**, that ethical considerations in engineering are sufficiently covered by safety considerations, is the argument that safety and ethics do not always map onto the same concerns. In the simple unsafe, unethical algorithm deployment example, we see a distinction: the safety analysis is concerned with the potential crash and the ethical analysis is concerned with questions of control authority. Both analyses evaluate contributing factors to some risk, but the mechanism of arriving at the risk is different. The difference in arrival mechanisms categorize the different types of concerns. In a system safety-type concern scenario, functional/hardware failures and software errors create *safety hazards* which contribute to risks. In an ethical-type concern scenario, unaddressed or improperly addressed moral questions and uncertainties create *ethical quandaries*¹⁵ which contribute to risks.

ethical concern: moral questions → *ethical quandaries* → risk

safety concern: failures/errors → *safety hazards* → risk

Sometimes those risks are eventually the same, if the ethical concern has safety consequences. Sometimes the risks are different, if the ethical concern does not necessarily pose a risk to physical safety. Consider this risk mechanism flow with the unsafe, unethical algorithm deployment example.

Example: Deployment of an algorithm that takes control away from the pilot and causes the aircraft to crash.

ethical concern

safety concern

ethical concern: questions and decisions about using automated systems for safety-critical functions → inaccurate pilot expectation of agency and control authority → displaced actual agency and control authority from pilot to machine resulting in authority struggle, irrecoverable loss of control, and ultimate crash

safety concern: failed sensor in single sensor architecture → unsafe flight control command → irrecoverable loss of control resulting in crash

This distinction provides the foundation for the ethical principles that should be supplemented to the OPs to create a set of OP²s that can provide safety and ethical assurance. The principles should provide a framework for systematically

¹⁴In the broad sense of "measurable," which does not require, but also does not necessarily exclude, the use of numbers. See [15], sense III, "To judge, assess, compare; to gauge against something."

¹⁵Note, this term is a first-pass use and has not been formalized for use in engineering in the same way that the term 'hazard' has been defined and adopted.

uncovering and suitably addressing the ethical quandaries that challenge the engineering of an acceptably ethical system. Now, we can look to these challenges to derive a candidate representative set of ethical assurance principles for developing software systems.

C. Deriving the Ethical Principles

While this exercise to derive ethical assurance principles for software systems is agnostic to the type of software system implementation, some subsets of algorithmic systems may provide inspirational content for the derivation. Many industries today are interested in a type of probabilistic automation systems which falls under the popular designation: Artificial Intelligence/Machine Learning (AI/ML)¹⁶. Many proposed and realized applications of AI/ML systems have either prompted discussion of or resulted in ethically questionable consequences [17]. As a result, an entire field of research is emerging on ethical principles to consider when developing AI systems [18]. We can leverage this research to narrow the possible ethical considerations into abstractions that have already been identified as important to consider during algorithmic system development.

The Alan Turing Institute distills much of the current research in the ‘Ethics of AI’ discourse within a paper called *Understanding Artificial Intelligence Ethics and Safety* [19]. The purpose of this paper is to “... provide end-to-end guidance on how to apply principles of AI ethics and safety to the design and implementation of algorithmic systems in the public sector” [19]. The author, Dr. David Leslie, defines AI ethics as a “... set of values, principles, and techniques that employ widely accepted standards of right and wrong to guide moral conduct in the development and use of AI technologies” [19]. These statements align to the foundational motivation behind our ethical assurance principles, so we will proceed by using the principles identified in this paper as content in our ethical principle derivation exercise¹⁷.

Leslie proposes four ethical principles to follow during the development and deployment of algorithmic systems: Fairness, Accountability, Sustainability, and Transparency. These principles are summarized in the paper with the bullets below:

- **Fairness:** All AI systems that process social or demographic data pertaining to features of human subjects must be designed to meet a minimum threshold of discriminatory non-harm. This entails that the datasets they use be equitable; that their model architectures only include reasonable features, processes, and analytic structures; that they do not have inequitable impact; and that they are implemented in an unbiased way.
- **Accountability:** All AI systems must be designed to facilitate end-to-end answerability and auditability. This requires both responsible humans-in-the-loop across the entire design and implementation chain and activity monitoring protocols that enable end-to-end oversight and review.
- **Sustainability:** Designers and users of AI systems must remain aware that these technologies have transformative effects on individuals and society. They must thereby proceed with a continuous sensitivity to real-world impacts. They must also keep in mind that the technical sustainability of these systems depends on their safety: their accuracy, reliability, security, and robustness.
- **Transparency:** Designers and implementers of AI systems must be able (1) to explain to affected stakeholders in everyday language how and why a model performed the way it did in a specific context and (2) to justify the ethical permissibility, the discriminatory non-harm, and the public trustworthiness both of its outcome and of the processes behind its design and use [19].

Leslie classifies these principles into two groups: normative¹⁸ criteria (fairness, sustainability) and procedural mechanisms (accountability, transparency). This notion of classifying the ethical principle as either normative or procedural is interesting because it relates to the subjective vs. objective conflict mentioned in Section III.A. Procedural methods are relatively easy to follow because they can be broken into discrete, defined steps. Normative judgments, however, rely on some consensus of what ought to be done, and thus can be more difficult to follow if the criteria to converge on consensus is not clear. Engineering as a discipline often attempts to remain fixed in the realm of objectivity, staying away from subjective judgments for fear of undermining the technical integrity of the system. Keeping these subjective elements out of the engineering process, however, may be creating an opportunity for ethical risk during deployment.

¹⁶See [16] for guidance on using more semantically accurate language to describe algorithmic systems without using anthropomorphizing language like AI/ML. While using the phrases ‘probabilistic automation’ and ‘model error reduction’ in place of ‘artificial intelligence’ and ‘machine learning’ may be more semantically accurate, the terms AI/ML are used in this section to match the reference material and avoid confusion.

¹⁷Using this research is just one way of deriving the ethical principles for the OP²s; completing further research and refinement which may alter the principles is entirely plausible. At this stage, the point is to demonstrate what the principles could look like, rather than propose rigid final principles.

¹⁸“In the context of practical ethics, the word ‘normativity’ means that a given concept, value, or belief puts a moral demand on one’s practices, i.e. that such a concept, value, belief indicates what one ‘should’ or ‘ought to’ do in circumstances where that concept, value, or belief applies” [19].

As such, applying normative ethics to the development of technology alone is quite difficult. Leslie's classification is clever considering this challenge. The normative principles (fairness, sustainability) are 'backed up' or reinforced by the procedural principles (accountability, transparency). First, the definitions of fair and sustainable are agreed upon. Then, the accountable agents are identified and the system is made transparent, both to systematically reinforce fairness and sustainability. This structural support built into the set of ethical principles may reasonably enable their adoption and inspire the derivation of overarching principles which can provide ethical assurance for a system.

The normative principles (fairness, sustainability) relate most closely to the **Intent** and **Innocuity** properties. Just as the **Intent** of a product is implied to be safe, so is the **Intent** implied to be ethical. Ensuring that this ethical **Intent** is realized and the system is fair and sustainable (and any other normative principle), then, encourages considering an explicit overarching ethical principle about what the system is supposed to do. It may add value to levy a general principle which aligns what the system is supposed to do with what the system ethically ought to do. This idea forms the first overarching principle: **Normative Alignment**. Generalizing this principle covers the definition of ethical values and principles in the conceptual stage, allows for the inclusion of normative principles like fairness, sustainability, and any other principles that are identified as ethically important, and ensures proper alignment before the **Intent** is written into the product as requirements that are then measured for **Innocuity** to provide grounds for safety assurance.

The procedural principles (accountability, transparency) relate most closely to the **Correctness** and **Innocuity** properties. **Correctness** requires the translucency, interpretability, and justifiability aspects of transparency to be imposed upon the **IMPLEMENTATION** to demonstrate that the product is an accurate manifestation of the design. Accountability, however, is not clearly covered by the OPs. The owner of the product has, or should have, no technical impact upon the safety of the product. In an ideal world, the **Innocuity** property will hold and the product will do no harm. In a practical world, **Innocuity** cannot be guaranteed. So, we may gain assurance value by interrupting this suspension of disbelief that the system will do no harm ever and introduce some coverage to handle the cases where that property fails. Thus, accountability is an important ethical feature of development as it provides coverage and reinforcement in the scenarios where the system may perform undesirably. This gap creates space for the second overarching principle: **Accountable Ownership**. The system itself is not eligible for moral accountability nor can it mitigate its own undesirable behavior; there must be a human agent who takes ownership of the system, understands the systems, and is responsible for the actions of that system.

D. Supplementing the Overarching Properties with Ethical Principles

So, we can use the normative criteria and procedural mechanism categories to reduce the set of four ethical principles into two Overarching Principles — **Normative Alignment** and **Accountable Ownership** — to supplement and bolster the Overarching Properties.

Overarching Properties

- a. **Intent**: The **DEFINED INTENDED BEHAVIOR** is correct and complete with respect to the **DESIRED BEHAVIOR**.
- b. **Correctness**: The **IMPLEMENTATION** is correct with respect to its **DEFINED INTENDED BEHAVIOR**, under **FORESEEABLE OPERATING CONDITIONS**.
- c. **Innocuity**: Any part of the **IMPLEMENTATION** that is not required by the **DEFINED INTENDED BEHAVIOR** has no **UNACCEPTABLE IMPACT**.

Overarching Principles

- a. **Normative Alignment**: The **DESIRED BEHAVIOR** is consistent with agreed-upon **NORMATIVELY ETHICAL BEHAVIOR**.
- b. **Accountable Ownership**: The **IMPLEMENTATION** is owned by human agents who are accountable for (1) **UNDERSTANDING** the part of the **IMPLEMENTATION** they own and (2) **TAKING RESPONSIBILITY FOR** the actions of that part of the **IMPLEMENTATION**.

Definitions

- a. **DESIRED BEHAVIOR**: Needs and constraints expressed by the stakeholders (this includes those needs and constraints identified by the **SAFETY ASSESSMENT**, the **ETHICS ASSESSMENT**, and those mandated by regulations).
- b. **DEFINED INTENDED BEHAVIOR**: The record of the **DESIRED BEHAVIOR**.
- c. **IMPLEMENTATION**: **ITEM** or combination of inter-related **ITEMS** for which acceptance or approval is being sought.
- d. **ITEM**: a hardware or software element having bounded and well-defined interfaces.
- e. **FORESEEABLE OPERATING CONDITIONS**: External and internal conditions in which the system is used, encompassing all known normal and abnormal conditions identified by the **SAFETY ASSESSMENT** and the **ETHICS ASSESSMENT**.

- f. **UNACCEPTABLE IMPACT**: An impact that compromises the **SAFETY ASSESSMENT** and/or the **ETHICS ASSESSMENT**.
- g. **SAFETY ASSESSMENT**: The systematic identification of **FAILURE CONDITIONS** and classifications in an operational context, evaluation of the architecture against safety objectives arising from these hazards, evaluation of potential common modes and threats, defining additional intended behaviors to support claims within these evaluations and showing that the safety objectives are satisfied by the **IMPLEMENTATION**.
- h. **FAILURE CONDITION**: “A condition having an effect on the aircraft and/or its occupants, either direct or consequential, which is caused or contributed to by one or more failures or errors, considering flight phase and relevant adverse operational or environmental conditions or external events.”
- i. **NORMATIVELY ETHICAL BEHAVIOR**: Needs and constraints explicitly deemed, by consensus, ethical with respect to all **STAKEHOLDERS**.
- j. **STAKEHOLDERS**: All people and organizations directly, **STAKEHOLDERS_d**, and indirectly, **STAKEHOLDERS_i**, associated with the system throughout its entire lifecycle, including but not limited to investors, designers, developers, approvers, operators, users, and/or bystanders, etc. Direct stakeholders, **STAKEHOLDERS_d**, must serve as moral representatives of those who cannot self-advocate, **STAKEHOLDERS_i**.
- k. **ETHICS ASSESSMENT**: The systematic identification and classification of **ETHICAL QUANDARIES** raised by the conceptualization of a system, the organizational conditions and practices within the design and development environment, the complexities of the deployment environment, and the consequential effects upon **STAKEHOLDERS** from operating the system. NOTE: This type of assessment is not yet formalized in engineering.
- l. **ETHICAL QUANDARY**: A state of uncertainty in determining the morally right thing to do in a certain situation which has a direct or consequential effect on the **STAKEHOLDERS**. NOTE: This notion is not yet formalized in engineering.
- m. **UNDERSTANDING**: The **STAKEHOLDERS** accountable for the **IMPLEMENTATION** know and can articulate how the **IMPLEMENTATION** (1) works, (2) meets the requirements, (3) may perform in undesirable ways identified by the **SAFETY ASSESSMENT** and/or the **ETHICS ASSESSMENT**, and (4) contains mitigations and controls to prevent the items identified by the **SAFETY ASSESSMENT** and the **ETHICS ASSESSMENT**.
- n. **TAKING RESPONSIBILITY FOR**: The **STAKEHOLDERS** accountable for the **IMPLEMENTATION** are morally and legally liable for and have the duty of mitigating any undesirable operational consequences of the **IMPLEMENTATION**.

Writing statements for **Normative Alignment** and **Accountable Ownership** prompts the definition of some additional nested ethical terms and notions. The definition of **NORMATIVELY ETHICAL BEHAVIOR** explicitly includes *all* **STAKEHOLDERS** in the ethical validation of the **Intent** of the system. The definition of **STAKEHOLDERS** explicitly includes all people and organizations who interact with the system in a direct or indirect capacity. Consensus on normatively ethical desired behavior for the system requires understanding what is ethical to all **STAKEHOLDERS**. Everyone must be able to agree on what the system should and should not do and these decisions must be documented with rationale.

This ethical consensus activity may be described by the notion of an **ETHICAL ASSESSMENT**. While this type of assessment is not yet formalized in engineering, an **ETHICAL ASSESSMENT** may be similar to that of a **SAFETY ASSESSMENT**. As a first pass, the definition of **ETHICAL ASSESSMENT** mirrors the **SAFETY ASSESSMENT** definition in establishing a systematic method for explicitly identifying and assessing the ethical contributions to some risk to avoid in the operation of the system. As mentioned in Section IV.B, the **SAFETY ASSESSMENT** evaluates failures which create safety hazards which contribute to risks. The **ETHICAL ASSESSMENT** evaluates moral uncertainties which create **ETHICAL QUANDARIES** which contribute to risks. This new term, **ETHICAL QUANDARY**, is also a first pass placeholder used to explicitly name the scenarios where moral uncertainties influence or undermine either the engineering process or the assessed consequence of the system. By identifying, evaluating, and controlling these scenarios in a systematic process, the **STAKEHOLDERS_d** can establish a more solid foundation of assurance that these scenarios will not pose latent risks to the system development and operation.

UNDERSTANDING and **TAKING RESPONSIBILITY FOR** are operative phrases in the **Accountable Ownership** statement, rather than terms with relatively static meanings. Nonetheless, the definitions of these phrases are provided as guidance for criteria to satisfy the spirit of each phrase. Setting criteria for these responsibilities is the most important feature for facilitating the demonstration of accountability. For example, ‘moral and legal liability’ is a strong burden to place on individuals in an accountability chain for a system and raises questions of external assignment and enforcement. Regardless of how the accountability principle is adhered to, when the system poses potentially severe safety and ethical consequences, responsibility must be defined or accountability will only be an illusion that is easily displaced between the many hands involved in the system. The action of assigning **Accountable Ownership** serves to remind all **STAKEHOLDERS_d** of the gravity of the situation and the potential effects on **STAKEHOLDERS_i**. The system should not be

approved and deployed until there is sufficient confidence that the system is safe and ethical enough for deployment. That is, the **Accountable Owners** have enough epistemic trust in the system that it is ready to be released without issue, and they know they will be responsible if there are any issues.

V. Ethical Principles in the Assurance Process

The **Normative Alignment** overarching ethical principle encourages the clear definition of the *STAKEHOLDERS* associated with the system, their objectives in developing the system, and the convergence upon an explicit determination that the objectives are ethical for *all STAKEHOLDERS* involved in the lifecycle of the system. This alignment makes the **Intent** property more robust as a measure of assurance that the system is suitable for approval. Initially confirming the ethics of the system provides stability and confidence in the foundational property of **Intent** that the following properties, **Correctness** and **Innocuity**, depend on. When developing assurance process artifacts for system approval, documenting activities to demonstrate adherence to the **Normative Alignment** principle results in the generation of additional pieces of evidence upon which to build justified confidence in the ethical **Intent** of the system.

The **Accountable Ownership** overarching ethical principle encourages the explicit determination of accountable human owners to systematically reinforce the assumed engineering knowledge of the system and the chain of accountability in situations where mitigations are required. This principle adds robustness to the **Correctness** and **Innocuity** properties in explicitly determining the individuals who are responsible for understanding the system thoroughly enough to be able to demonstrate its **Correctness** and **Innocuity**, and identifying the individuals who will take accountability for any unintended consequences of deploying and operating the system. These actions already exist in the engineering process and are made explicit by this principle for the purpose of supplementing the assurance process for the system. This explicit, proactive identification does, however, change the mitigation process by shifting forward incident reaction time. System owners and approvers would no longer require time following an incident to identify the people responsible for mitigating the consequence, instead they could swiftly enact the predetermined mitigation plans prompted by the principle. These plans and the documentation of the activities and decisions to demonstrate adherence to the **Accountable Ownership** principle results in the generation of additional pieces of evidence upon which to build justified confidence in the ethical approval and deployment of the system.

As principles to follow during the entire lifecycle of the system, **Normative Alignment** and **Accountable Ownership** also provide opportunities to continually monitor and assess the ethical and safety implications of the system. If operational context changes, these steps can be performed again to ensure that the system is suitably ethical for all *STAKEHOLDERS* and can serve as catalysts to iterate on the assurance process if necessary to proactively make system updates to reassure confidence in the system. If the system performs in an undesirable manner, the principles speed up reaction time to perform mitigation steps in an efficient and effective way.

A. Hypothesis Findings

These overarching ethical principles derived from an informal ethical analysis on the OPs theory which used a thought experiment to explore the question: (1) do the OPs ensure adequate coverage of ethics? To properly answer that question, some more questions were asked: (2) why are ethics not explicit systematic considerations in engineering? and (3) should we make them explicit? This line of thinking brought us to some hypotheses about ethics in engineering which provided a grounding point to use as the center of the ethical analysis of the OPs. In evaluating the OPs for implied ethical coverage, we can identify existing support for the hypotheses. In analyzing these supporting notions, we can scrutinize the practice of solely implying ethics in engineering and propose a rejection of relying on ethical implications in engineering alone.

Ethics are not explicitly considered as a discipline in engineering because...

Hypothesis A: Considering safety in engineering is sufficient for considering ethics.

Existing support: The desired behavior for the system defined by the stakeholders includes those needs and constraints identified by the safety assessment, which implies coverage of ethical concerns with safety consequences. The innocuous property of the system assures that the system will do no wrong, which implies no ethically wrong things. The safety assessment could be extended to include ethical harms.

Reject because: Safety and ethics do not always map onto the same type of concerns. We cannot be sure that the safety assessment process *as it is practiced today* is robust to uncovering all potential causes

of ethical concerns which pose a risk of unethical deployment, in addition to safety concerns. We also cannot guarantee that all ethical concerns will map to safety consequences.

Hypothesis B: Ethics are covered by the good intentions of the product developers and societal norms reinforced by regulations.

Existing support: The desired behavior for the system defined by the stakeholders is implied to be naturally ethical by good intentions or otherwise ethical by regulation. Reactive regulations have been adopted to mandate protections against specific socially unacceptable events.

Reject because: Trusting in good intentions and societal norms is neither systematically defensible nor creates engineering artifacts to support an assurance argument. By comparison, making safety considerations explicit in engineering made systems better without invalidating trust in the good intentions and societal norms for suitable technologies. Engineers have a moral duty to determine if making ethical considerations explicit will make systems better.

The hypotheses and their existing support answer the question: why are ethics not explicit systemic considerations in engineering? Imagine we use our thought experiment to reject these hypotheses. Then we have the answer to the third question — should we make ethics explicit? — in the affirmative. Yes, we should make ethics explicit in engineering.

B. OP²s Findings

This affirmation leads to a few more questions: (4) what might we glean by directing resources to focus on the ethical consequence not just as a contributing factor to a safety consequence but as a locus of control for assessment itself? (5) would it be valuable to generate ethical assurance artifacts for complex systems, in addition to safety assurance artifacts? and (6) how would we generate those assurance artifacts? In our thought experiment, these questions are answered by the identification of ethical engineering concerns and quandaries that are independent to items that serve as the focus of the safety assessment and the derivation of a method, the OP²s, to uncover and systematically address them. Consider an OP²s assurance framework for the running example below:

We can have sufficient assurance in the suitability of an example automated system that has the ability to perform safety-critical flight control functions with the following OP²s framework:

Intent: A safe, ethical flight control system algorithm that assists the pilot by automatically handling flying quality adjustments.

Correctness: The algorithm has the appropriate inputs, data processing commands, and outputs to accurately perform only the required flying quality adjustments.

Innocuity: The algorithm has the appropriate sensor and system redundancies, fault tolerant architecture, and control authority. The algorithm has no other functional capabilities.

Normative Alignment: The intention for the algorithm to assist the pilot to improve task management is agreed upon as acceptably ethical.

Example Rationale: The stakeholders should evaluate all hardware and software solutions to the flying qualities engineering problem before deciding on a software solution. The agent with current control authority should be clear at all times. In an emergency, the pilot should always be able to disengage the software and take manual flight controls without authority struggle. The pilot should be informed of the existence of the algorithm and the functional capability of the algorithm. The pilot should be trained to handle failures of system. The passengers should be informed that the pilot is operating the aircraft in conjunction with some automated flight controls.

Accountable Ownership: The accountable owners and their responsibilities are identified for each part of the implementation.

Example Rationale: The accountable owners should understand and be able to explain the part of the implementation they are responsible for. The accountable owners should know and be able to explain the ethical and safety permissibility of their part of the implementation. The accountable owners should know the exact boundaries of their part of the implementation, the requirements and analyses associated

with their part of the implementation, and the purpose and function of each line and/or block of code in their part of the implementation. The accountable owners should be able to explain in everyday language how and why the implementation performs in specific ways. The accountable owners should know that they are responsible for publicly deploying the implementation and taking moral and legal accountability if the implementation performs undesirably. The accountable owners should create and be prepared to implement mitigation plans for contingency and incident management.

In theory, the first three properties provide grounds for sufficient epistemic confidence — *assurance* — that the system is suitably engineered. In practice, the next two principles provide additional grounds for sufficient epistemic confidence — *reassurance* — that the system is suitable for public deployment. This nuanced observation provides an answer to the original question, does the sufficient set of OPs ensure adequate coverage of *ethics*? In other words, does demonstrating the possession of the three OPs in a system assure stakeholders that the product will be *ethical enough* for public deployment? Sometimes yes. Sometimes no.

Historically, we have observed cases where systems have been designed and developed according to the assurance process, approved, deployed, and still performed in unacceptable ways; for two examples, see [20, 21]. Now, with increasingly automated systems performing functions that have safety and ethical consequences, we need increasingly strong arguments to justify confidence in the deployment of those systems. In the cases where ethical consequences are potentially severe, relying on implied ethics in the engineering process does not result in adequate ethical assurances for suitable deployment.

The ethical principle statements and rationale in the OP²s framework example above provide some insight into the types of questions and decisions that would be explicitly considered with documentation during the assurance process if something like the OP²s was adopted. Each item in the rationale would catalyze several engineering processes, resulting in a robust collection of assurance artifacts. Thus, we conclude our thought experiment by suggesting that in cases with potentially severe ethical implications, ethics should be explicitly considered in engineering. The OP²s extension is just an example of one way to systematically perform this explicit consideration.

VI. Concluding Remarks

In a world where autonomy increasingly shifts from the human to the machine, the risk of unintended consequences also increases [5, 22]. These complex, increasingly autonomous systems have potentially severe safety and ethical consequences, so they require appropriate safety *and* ethical assurances before approving for public deployment. Introducing a framework for ethical analysis, like the OP²s, can add value to the assurance process by systematically generating artifacts to demonstrate this expanded set of assurances.

Beyond just existing in engineering, ethical considerations are *fundamental* to engineering. Making these fundamental ethical considerations explicit may be the key to managing those unintended consequences. **Normative Alignment** makes explicit the ethical implications assumed within the **Intent** of the system and **Accountable Ownership** makes explicit the ethical implications assumed within the development of a **Correct** and **Innocuous** system. Together, the overarching set of properties and principles proposed by the OP²s provides a systematic way to make explicit the assumptions regarding the ethics and safety of the system in service of generating assurances that the system will be both ethical and safe enough for approval and deployment.

As with the OPs, the means of demonstrating the OP²s are not prescriptive. These ideas are proposed to describe the theoretical intent behind the assurance processes. Once the theory is well established, then new methods for demonstrating the possession of overarching assurance properties and the adherence to overarching assurance principles can be defined. This first principles approach facilitates flexibility and creativity in generating artifacts for system assurance arguments while also expanding the coverage of assurances. Introducing the explicit consideration of ethics as a systematic process in engineering also paves the way for a discipline to emerge and evolve which will define guidelines for determining right and wrong engineering processes and consequences. Careful and meticulous consideration of ethics can only improve the suitability of systems being deployed in our current and future complex, interconnected technological landscapes.

Acknowledgments

This work was completed during the Spring 2026 session of the NASA Langley Research Center Office of Science, Technology, Engineering, and Mathematics (OSTEM) Internship program with the expert guidance of C. Michael Holloway. Foundational ideas for this work were developed in part during my Master of Philosophy degree in the Ethics of Artificial Intelligence, Data, and Algorithms at the University of Cambridge, with academic guidance from Claire Benn, Ph.D. and John Burden, Ph.D.

These ideas build upon the work of all the people involved in the Overarching Properties Working Group from 2015-2019 and all the OPs work that came after to refine the theory. A special thanks goes to Lauren White, Ph.D., Mallory Graydon, Ph.D., and Aaron Dutle, Ph.D. who provided invaluable guidance, brainstorming advice, and review.

To Reach the Author

Send e-mail to taylor.thoman@nasa.gov and c.michael.holloway@nasa.gov.

References

- [1] Holloway, C. M., “Understanding the Overarching Properties,” Technical Memorandum NASA/TM-2019-219650, National Aeronautics and Space Administration, Hampton, VA, USA, July 2019. URL <https://hdl.handle.net/2060/20190029284>.
- [2] Britannica Editors, “normative ethics,” Encyclopedia Britannica, June 2023. URL <https://www.britannica.com/topic/normative-ethics>.
- [3] Luettig, B., Akhiat, Y., and Daw, Z., “ML meets aerospace: challenges of certifying airborne AI,” *Frontiers in Aerospace Engineering*, Vol. Volume 3 - 2024, 2024. <https://doi.org/10.3389/fpace.2024.1475139>, URL <https://www.frontiersin.org/journals/aerospace-engineering/articles/10.3389/fpace.2024.1475139>.
- [4] Koopman, P., and Widen, W., “Redefining safety for autonomous vehicles,” *International Conference on Computer Safety, Reliability, and Security*, Springer, 2024, pp. 300–314. URL <https://arxiv.org/abs/2404.16768>.
- [5] McCormick, F., Graydon, M., Neogi, N., Miner, P., and Maddalon, J., “Safety Expertise and the Perils of Novelty,” *2023 IEEE/AIAA 42nd Digital Avionics Systems Conference (DASC)*, IEEE, 2023, pp. 1–10.
- [6] Wasson, K. S., and Voros, R., “Deobfuscating Machine Learning Assurance and Approval,” *2024 AIAA DATC/IEEE 43rd Digital Avionics Systems Conference (DASC)*, 2024, pp. 1–10. URL <https://api.semanticscholar.org/CorpusID:274061124>.
- [7] Paul, S., Prince, D., Iyer, N., Tang, L., Durling, M., Meng, B., Meiners, M., and Varanasi, S. C., “Overarching Properties-Based Arguments for the Safety Assurance of an Ai-Assisted Runway Alignment System,” *2025 AIAA DATC/IEEE 44th Digital Avionics Systems Conference (DASC)*, 2025, pp. 1–10. URL <https://api.semanticscholar.org/CorpusID:283475711>.
- [8] Paul, S., Iyer, N., Prince, D., Tang, L., Durling, M., Meiners, M., Meng, B., Visnevski, N., and Mandal, U., “Assurance of AI/ML-Based Aerospace Systems Using Overarching Properties,” *2024 AIAA DATC/IEEE 43rd Digital Avionics Systems Conference (DASC)*, 2024, pp. 1–10. <https://doi.org/10.1109/DASC62030.2024.10749102>.
- [9] Daw, Z., and Holloway, C. M., “From Skepticism to Understanding: Correcting Misconceptions About Overarching Properties,” *2025 IEEE/AIAA 44th Digital Avionics Systems Conference (DASC)*, 2025. URL <https://ntrs.nasa.gov/citations/20250006016>.
- [10] Daw, Z., Beecher, S., and Holloway, M., “Leveling Arguments: Easier Said than Done,” *2023 IEEE/AIAA 42nd Digital Avionics Systems Conference (DASC)*, Barcelona, Spain, 2023. URL <https://ntrs.nasa.gov/citations/20230009081>.
- [11] Daw, Z., and Beecher, S., “Assuring safety in a flexible aerospace certification — Lessons learned on applying OPs at the system level—,” *2023 IEEE International Systems Conference (SysCon)*, 2023, pp. 1–8. <https://doi.org/10.1109/SysCon53073.2023.10131225>.
- [12] Toulmin, S. E., *The Uses of Argument*, updated edition (1958 original) ed., Cambridge University Press, Cambridge, UK, 2003.
- [13] Govier, T., *A Practical Study of Argument*, 7th ed., Wadsworth/Cengage Learning, 2014.
- [14] Oxford University Press, “epistemic, adj.” Oxford English Dictionary, September 2025. URL <https://doi.org/10.1093/OED/4330537275>.
- [15] Oxford University Press, “measure, v.” Oxford English Dictionary, March 2026. URL <https://doi.org/10.1093/OED/7185392046>.

- [16] Inie, N., Zukerman, P., and Bender, E. M., “De-anthropomorphizing “AI”: From wishful mnemonics to accurate nomenclature,” *First Monday*, 2026.
- [17] Ghosh, M., “Artificial intelligence (AI) and ethical concerns: a review and research agenda,” *Cogent Business & Management*, Vol. 12, No. 1, 2025, p. 2551809. <https://doi.org/10.1080/23311975.2025.2551809>.
- [18] Kazim, E., and Koshiyama, A. S., “A high-level overview of AI ethics,” *Patterns*, Vol. 2, No. 9, 2021.
- [19] Leslie, D., “Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector,” Tech. rep., The Alan Turing Institute, 2019. <https://doi.org/10.5281/ZENODO.3240529>, URL <https://zenodo.org/record/3240529>.
- [20] Australian Transportation Safety Bureau, “In-flight upset event 240km north-west of Perth, WA Boeing Company 777-200, 9M-MRG, 1 August 2005,” Tech. rep., Aviation Occurrence Report AOR-200503722 Final, 13 March 2007.
- [21] Federal Aviation Administration, “Summary of the FAA’s Review of the Boeing 737 MAX: Return to Service of the Boeing 737 MAX Aircraft,” U.S. Department of Education, November 2020. URL https://www.faa.gov/sites/faa.gov/files/2022-08/737_RTS_Summary.pdf.
- [22] Cofer, D., “Unintended behavior in learning-enabled systems: detecting the unknown unknowns,” *IEEE/AIAA 40th Digital Avionics Systems Conference (DASC)*, 2021, pp. 1–7.