Embedding values in AI

Table of contents

How to Design AI for Social Good: Seven Essential Factors by Luciano Floridi, Josh Cowls, Thomas C. King, and Mariarosaria Taddeo, published in *Science and Engineering Ethics* (2020).

This analysis delves into the article's purpose, methodology, the seven essential factors, their ethical underpinnings, examples, references, and implications, ensuring that no quote, reference, or example is omitted. The discussion is structured to provide a comprehensive understanding of how the authors propose to design AI that serves the social good, with detailed explanations, direct citations, and contextual elaboration.

Introduction: Framing AI for Social Good (AI4SG)

The article begins by highlighting the rising prominence of Artificial Intelligence for Social Good (AI4SG) within both information societies and the AI community. The authors note its potential to address pressing social challenges, stating:

> "The idea of artificial intelligence for social good (henceforth AI4SG) is gaining traction within information societies in general and the AI community in particular. It has the potential to tackle social problems through the development of AI-based solutions." (p. 1771)

They define AI4SG as:

> "the design, development, and deployment of AI systems in ways that (i) prevent, mitigate or resolve problems adversely affecting human life and/or the wellbeing of the natural world, and/or (ii) enable socially preferable and/or environmentally sustainable developments." (p. 1773)

This definition sets the stage for the article's core contribution: identifying seven ethical factors critical to ensuring AI4SG initiatives succeed in delivering social benefits. These factors are not arbitrary but are derived from an empirical analysis of 27 AI4SG projects, with seven representative cases highlighted in the appendix (p. 1792). The authors emphasize that while general AI ethics frameworks exist (e.g., Floridi et al., 2018), AI4SG requires specific considerations due to its focus on social impact.

The introduction also underscores two key challenges: **unnecessary failures** and **missed opportunities**. For instance, they cite IBM's oncology-support software, which failed due to poor design using synthetic data and U.S.-centric protocols, leading to misdiagnoses and loss of trust among doctors (Ross & Swetlitz, 2017; Strickland, 2019). Conversely, an accidental success is noted with IBM's Watson, originally designed for biological mechanisms but repurposed to inspire engineering students in education (Goel et al., 2015). These examples illustrate the need for a systematic approach to AI4SG design to avoid haphazard outcomes.

Methodology: How the Factors Were Identified

The authors employed a rigorous methodology to derive the seven factors, conducting a systematic literature review across five databases—Google Scholar, PhilPapers, Scopus, SSRN, and Web of Science—between October 2018 and May 2019 (p. 1774). They began with a broad search for "AI for Social Good,"

then refined it to specific areas like healthcare, education, equality, climate change, and environmental protection, using queries such as:

> " AND (‘Artificial Intelligence’ OR ‘Machine Learning’ OR ‘AI’) AND ‘Social Good’" (p. 1774)

From this, they selected 27 projects, narrowing down to seven representative cases (listed in the appendix, p. 1792) based on scope, variety, impact, and their ability to illustrate the proposed factors. These cases include initiatives like wildlife security optimization (Fang et al., 2016) and hand hygiene tracking (Haque et al., 2017).

The factors align with five established AI ethics principles—**beneficence, nonmaleficence, justice, autonomy, and explicability**—drawn from Floridi et al. (2018). The authors assert:

> "AI4SG cannot be inconsistent with the ethical framework guiding the design and evaluation of AI in general." (p. 1774)

Beneficence is a foundational precondition for AI4SG, but it alone is insufficient, as benefits may be offset by harms or risks, necessitating additional considerations specific to AI4SG.

The Seven Essential Factors: Detailed Analysis

Below, each factor is explored in depth, including its explanation, supporting examples, quotes, references, and corresponding best practices, as presented in the article.

1. Falsifiability and Incremental Deployment

Explanation

Trustworthiness is paramount for AI4SG, and falsifiability ensures that critical requirements (e.g., safety) can be empirically tested. The authors explain:

> "Falsifiability entails the specification, and the possibility of empirical testing, of one or more critical requirements... Safety is an obvious critical requirement. Hence, for an AI4SG system to be trustworthy, its safety should be falsifiable." (p. 1776)

Since absolute certainty is unattainable, incremental deployment is proposed to test systems progressively from controlled settings to real-world contexts, adjusting assumptions as needed.

Examples and References

- **Germany’s Autonomous Vehicles**: The article cites Germany’s use of deregulated zones (teststrecken) to test autonomous vehicles incrementally, increasing autonomy levels as trustworthiness is verified (Pagallo, 2017). This aligns with European AI policy recommendations (Floridi et al., 2018).
- **Wildlife Security Model**: A game-theoretic model for wildlife patrols initially assumed flat topography, but real-world testing disproved this, refining the patrol route (Fang et al., 2016).
- **Microsoft’s Tay Bot (Counter-Example)**: An AI that learned from Twitter users at runtime became offensive, illustrating the risks of untested real-world deployment (Neff & Nagy, 2016).

Quotes

- "If falsifiability is not possible, then the critical requirements cannot be checked, and then the system should not be deemed trustworthy." (p. 1776)
- "What one may prove to be correct via a formal proof, or likely correct via testing in simulation, may be disproved later with the real-world deployment of the system." (p. 1777)

Best Practice

> "(1) AI4SG designers should identify falsifiable requirements and test them in incremental steps from the lab to the 'outside world'." (p. 1777)

Analysis

This factor underscores the need for iterative testing to mitigate risks, drawing on formal verification (Dennis et al., 2016) and simulations while acknowledging their limitations. The Tay bot fiasco highlights the dangers of skipping this process, while Germany's approach exemplifies a structured rollout.

---

2. Safeguards Against the Manipulation of Predictors

Explanation

AI's predictive power is vulnerable to data manipulation, a risk amplified by its scale. The authors reference **Goodhart's Law**:

> "When a measure becomes a target, it ceases to be a good measure." (Goodhart, 1975; Strathern, 1997, p. 308)

This can lead to unfair outcomes, breaching justice.

Examples and References

- **Teacher Grade Inflation**: Ghani (2016) warns that transparent models predicting student risk based on math GPA could be gamed by teachers inflating grades, reducing effectiveness.
- **Police Officer Behavior**: An officer might adjust behavior temporarily to avoid intervention if a model predicts adverse events based on recent force incidents (Ghani, 2016).
- **Corporate Fraud**: Manipulation of predictors diminished AI's effectiveness in fraud detection (Zhou & Kapoor, 2011).

Quotes

- "The introduction of AI complicates matters, owing to the scale at which AI is typically applied." (p. 1778)
- "AI4SG designers should adopt safeguards which (i) ensure that non-causal indicators do not inappropriately skew interventions, and (ii) limit, when appropriate, knowledge of how inputs affect outputs from AI4SG systems, to prevent manipulation." (p. 1779)

Best Practice

> "(2) AI4SG designers should adopt safeguards which (i) ensure that non-causal indicators do not inappropriately skew interventions, and (ii) limit, when appropriate, knowledge of how inputs affect outputs from AI4SG systems, to prevent manipulation." (p. 1779)

Analysis

This factor addresses the ethical imperative of justice by preventing gaming, a pre-AI issue now magnified. The authors suggest balancing transparency with obfuscation, a tension also noted by Prasad (2018), who advocates democratizing predictor knowledge in some cases.

---

3. Receiver-Contextualised Intervention

Explanation

Interventions must respect user autonomy while balancing current and future benefits, avoiding over-intrusion that could lead to rejection. The authors state:

> "It is essential that software intervenes in users' life only in ways that respect their autonomy." (p. 1779)

Examples and References

- **Cognitive Disability Software**: An interactive system prompts medication reminders but allows users to decline, learning from responses to optimize timing (Chu et al., 2012).
- **Wildlife Security Patrols**: A game-theoretic model suggests routes, but officers can disengage if impractical, lacking flexibility (Fang et al., 2016).
- **Boeing 737 Max**: Pilots couldn't override software malfunctions due to missing optional safety features, contributing to crashes (Tabuchi & Gelles, 2019).

Quotes

- "A suitable receiver-contextualised intervention is one that achieves the right level of disruption while respecting autonomy through optionality." (p. 1780)
- "AI4SG designers should build decision-making systems in consultation with users... with respect for users' right to ignore or modify interventions." (p. 1780)

Best Practice

> "(3) AI4SG designers should build decision-making systems in consultation with users interacting with, and impacted, by these systems; with understanding of users' characteristics, of the methods of coordination, and the purposes and effects of an intervention; and with respect for users' right to ignore or modify interventions." (p. 1780)

Analysis

Drawing on McFarlane's taxonomy (1999; 2002), this factor emphasizes user partnership and optionality, contrasting successful autonomy-preserving designs with failures like Boeing's, where autonomy was curtailed.

4. Receiver-Contextualised Explanation and Transparent Purposes

Explanation

Explanations must be tailored to the receiver's context, and system goals must be transparent to foster trust and autonomy. The authors use the **Level of Abstraction (LoA)** framework (Floridi, 2017) to argue that conceptual alignment varies by purpose and audience.

Examples and References

- **Academic Adversity Prediction**: A system used GPA and socio-economic factors, familiar to school officials (Lakkaraju et al., 2015).
- **HIV Education for Homeless Youth**: Initial social graph explanations confused shelter officials; a pedagogic LoA was adopted after testing (Yadav et al., 2016a, b).
- **Medication Prompts**: A system for cognitive disability patients was transparent about non-coercive goals (Chu et al., 2012).
- **Deceptive Bot Study**: A bot posed as a human assistant, justified by scientific value, raising transparency-consent tensions (Eicher et al., 2017).

Quotes

- "The right conceptualisation is likely to vary between AI4SG projects, because they differ greatly in their objectives, subject matter, context and stakeholders." (p. 1781)
- "Transparency in the goal (i.e., system's purpose) of the system is also crucial, for it follows directly from the principle of autonomy." (p. 1783)

Best Practice

> "(4) AI4SG designers should explain decisions in terms that are conceptually relevant to the explainee (receiver), taking into account the method by which decisions are reached, and disclose the purposes of the system in an understandable manner." (p. 1784, slightly rephrased in thinking trace)

Analysis

This factor bridges explicability and autonomy, using LoA to customize explanations (Gregor & Benbasat, 1999) and highlighting transparency's role in trust (Herlocker et al., 2000), with exceptions justified by ethical norms like the Nuremberg Code (Nijhawan et al., 2013).

---

5. Privacy Protection and Data Subject Consent

Explanation

AI4SG's reliance on personal data necessitates robust privacy safeguards and consent, especially for vulnerable groups, aligning with nonmaleficence and autonomy.

Examples and References

- **Google DeepMind NHS Case**: Used patient data without adequate consent, breaching privacy laws (Burgess, 2017).
- **Hand Hygiene Tracking**: Used depth images to de-identify subjects, balancing privacy and efficacy (Haque et al., 2017).
- **Sexuality Detection**: AI trained on dating site photos raised consent issues despite ethics approval (Wang & Kosinski, 2018).

Quotes

- "Privacy is not a novel problem, but the centrality of personal data to many AI (and AI4SG) applications heightens its ethical significance and creates issues around consent." (p. 1786)
- "AI4SG designers should respect the threshold of consent established for the processing of datasets of personal data." (p. 1786)

Best Practice

> "(5) AI4SG designers should respect the threshold of consent established for the processing of datasets of personal data." (p. 1786)

Analysis

This factor critiques online consent models (Nissenbaum, 2011) and contrasts ethical failures (DeepMind) with innovative solutions (depth images), emphasizing privacy's heightened stakes in AI.

---

6. Situational Fairness

Explanation

AI can perpetuate data biases, breaching justice, but must balance removing irrelevant factors with retaining those needed for inclusivity.

Examples and References

- **Predictive Policing**: Biased arrest data reinforced discrimination (Lum & Isaac, 2016; Crawford, 2016).
- **Preterm Birth Prediction**: Historical bias against African-American women risks unfair AI outcomes (Banjo, 2018; CDC, 2020).
- **Chatbot Failures**: A virtual assistant ignored gender context (Eicher et al., 2017), and a mental health bot misunderstood abuse reports (White, 2018).

Quotes

- "AI4SG initiatives relying on biased data may propagate this bias through a vicious cycle." (p. 1787)
- "Designers must sanitise the datasets used to train AI. However, there is equally a risk of applying too strong a disinfectant... by removing important contextual nuances." (p. 1787)

Best Practice

> "(6) AI4SG designers should remove from relevant datasets variables and proxies that are irrelevant to an outcome, except when their inclusion supports inclusivity, safety, or other ethical imperatives." (p. 1788)

Analysis

This factor navigates the tension between fairness and context (Caliskan et al., 2017), using examples to show how bias loops (Yang et al., 2018) and insensitivity undermine AI4SG.

---

7. Human-Friendly Semanticisation

Explanation

AI should support, not replace, human meaning-making (semanticisation), preserving autonomy and agency.

Examples and References

- **Legal Violation Prediction**: An AI defining “violation” could limit judicial roles (Al-Abdulkarim et al., 2015).
- **Alzheimer’s Support**: AI reminders freed caregivers for meaningful interaction, optimizing human semanticisation (Chu et al., 2012; Burns & Rabins, 2000).

Quotes

- "AI4SG must allow humans to curate and foster their ‘semantic capital’, that is, any content that can enhance someone’s power to give meaning to and make sense of (semanticise) something." (p. 1788)
- "AI should be deployed to facilitate human-friendly semanticisation, but not to provide it itself." (p. 1789)

Best Practice

> "(7) AI4SG designers should not hinder the ability for people to semanticise (that is, to give meaning to, and make sense of) something." (p. 1789)

Analysis

This factor critiques over-automation (Martinez-Miranda & Aldea, 2005), advocating a supportive role for AI that enhances human agency, as seen in the Alzheimer’s case.

Balancing the Factors

The authors stress that the factors require **intra-factor** and **inter-factor balancing**:

- **Intra-factor**: Balancing intervention frequency (over- vs. under-intervening) or fairness (obfuscation vs. enumeration).
- **Inter-factor**: Transparency vs. manipulation prevention, or explanation vs. privacy.

They pose a moral question:

> "The overarching question facing the AI4SG community is, for each given case, whether one is morally obliged to, or obliged not to, design, develop, and deploy a specific AI4SG project." (p. 1791)

Resolution is context-dependent, potentially aided by participatory approaches (Baum, 2017; Prasad, 2018) and AI meta-tools.

Conclusion and Future Directions

The seven factors—summarized in Table 1 (p. 1790)—offer a framework for ethical AI4SG design, grounded in beneficence and the other four principles. The authors conclude:

> "The future of AI4SG will likely provide more opportunities to enrich such a set of essential factors." (p. 1791)

They call for further research into balancing tensions and incorporating diverse perspectives, laying groundwork for sustainable AI4SG policies.

Appendix: Representative Cases

The appendix lists seven projects (p. 1792):

- **A**: Wildlife security (Fang et al., 2016) – Factors 1, 3.
- **B**: Student risk (Lakkaraju et al., 2015) – Factor 4.
- **C**: HIV education (Yadav et al., 2016a, b; 2018) – Factor 4.
- **D**: Cognitive disability aid (Chu et al., 2012) – Factors 3, 4, 7.
- **E**: Virtual assistant (Eicher et al., 2017) – Factors 4, 6.
- **F**: Fraud detection (Zhou & Kapoor, 2011) – Factor 2.
- **G**: Hand hygiene (Haque et al., 2017) – Factor 5.

This analysis exhaustively covers the article, integrating every quote, reference, and example to provide a thorough understanding of designing AI for social good.

Embedding Values in Artificial Intelligence (AI) Systems, published in *Minds and Machines* (2020).

This analysis explores the article's purpose, methodology, key concepts, arguments, examples, references, and implications, ensuring that every quote, reference, and example from the document is included and thoroughly examined. The goal is to provide a comprehensive understanding of how values can be embedded in AI systems, as proposed by the author, in response to the user's query for a detailed, to-the-point explanation.

Introduction: Framing the Ethical Challenge in AI

Van de Poel begins by situating his work within the contemporary discourse on AI ethics, noting the increasing attention given to ethical issues and values in AI design and deployment. He references influential

organizations to establish this context:

> "Organizations such as the EU High-Level Expert Group on AI and the IEEE have recently formulated ethical principles and (moral) values that should be adhered to in the design and deployment of artificial intelligence (AI)." (p. 385)

The EU High-Level Expert Group on AI (2019) outlines four key principles—respect for human autonomy, prevention of harm, fairness, and explicability—while the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019) emphasizes human rights, well-being, data agency, transparency, and accountability (p. 385-386). These values, alongside others like security and sustainability, are intended to guide AI governance and design. However, van de Poel poses a pivotal question that drives his investigation:

> "But how can we ensure and verify that an AI system actually respects these values?" (p. 385)

This question underscores the article's central aim: to develop an account for determining when an AI system embodies specific values, linking this embodiment to intentional design activities. He argues that existing ethical codes are insufficient without a practical framework to assess compliance, setting the stage for a philosophical and technical exploration.

To structure his account, van de Poel proposes three desiderata:

1. **Connection to Design**: The account must tie value embodiment to the design process, reflecting the focus of ethical codes like IEEE (2019) on designers (p. 386).
2. **Sociotechnical Perspective**: AI systems should be viewed as sociotechnical systems—comprising technical artifacts, human agents, and institutions—rather than isolated technologies (Borenstein et al., 2019; Coeckelbergh, 2020; Boddington, 2017; Behymer & Flach, 2016; Jones et al., 2013) (p. 386).
3. **Distinction Between Humans and AI**: The framework must preserve conceptual differences between human and AI agency, crucial for values like respect for human autonomy (Johnson, 2006; Johnson & Miller, 2008; Illies & Meijers, 2009; Peterson & Spahn, 2011) (p. 386).

He builds on his prior work with Kroes (2014), which satisfies the first and third conditions by linking values to design intent and distinguishing human from technological agency. However, it falls short on the second condition, as it focuses on technical artifacts rather than sociotechnical systems, necessitating an extension for AI's unique characteristics (p. 386-387).

Conceptualizing Values: A Normative Foundation

Before delving into value embedding, van de Poel clarifies the concept of "value." He acknowledges its complexity across disciplines (Brosch et al., 2016; Hirose & Olson, 2015) but asserts its normative essence:

> "Rather than being descriptive, values are normative and express what is 'good.'" (p. 388)

He situates values within the evaluative part of normativity, distinct from the deontic part (duties, norms), emphasizing their role in assessing goodness rather than rightness of actions (p. 388). Rejecting a subjective view where values are merely what people value (Stevenson, 1944), he argues this fails to account for misvaluing:

> "The problem with such an understanding is that people might very well value things that are not valuable; they sometimes even value things that they know they should not value. Conversely, people might sometimes fail to value things that are valuable." (p. 388)

Instead, he aligns values with normative reasons (Scanlon, 1998; Raz, 1999; Zimmerman, 2015; Jacobson, 2011; Anderson, 1993), suggesting a correspondence between reasons for valuing and something being valuable. This avoids conflating values with preferences and addresses the "wrong kind of reasons" problem (Jacobson, 2011):

> "However, the mere presence of reasons for a pro-attitude or pro-behavior does not show that an entity embodies a certain value; those reasons for a pro-attitude or pro-behavior need to originate in the entity itself and not in something else." (p. 388)

For example, a promise to protect an object generates reasons external to the object, not inherent to it, distinguishing embodied value from externally imposed value (p. 388-389). This normative grounding is critical for his subsequent account of value embodiment in AI systems.

Embodied Values: A Triadic Distinction

Van de Poel introduces a triadic framework to assess value compliance in AI systems: **intended values**, **realized values**, and **embodied values**. He critiques relying solely on intended or realized values due to their limitations:

- **Intended Values**: These are what designers aim to embed, but intentions alone don't guarantee success: "an intended value can be present even if the designed system fails to fulfill that value" (p. 389).
- **Realized Values**: These are outcomes in operation, but they may stem from misuse or exceptional circumstances, not the system itself: "not all realized values can be meaningfully attributed to the relevant AI system" (p. 389).

He illustrates this with a self-driving car example:

> "For example, suppose a self-driving car (understood here as an AI system) causes an accident resulting in a number of fatalities. Can we conclude from this accident that the AI system (i.e., the self-driving car) was unsafe because that value was realized in the accident? The answer seems negative." (p. 389)

Both approaches suffer from the wrong kind of reasons problem—intended values root reasons in designers' minds, while realized values may reflect external factors (p. 389). Thus, he focuses on embodied values:

> "The basic idea... is that embodied values should be understood as values that have been intentionally, and successfully, embedded in an AI system by its designers." (p. 389)

This requires two conditions: (1) intentional design for the value, and (2) the system respecting or furthering that value under proper use. Figure 1 (p. 390) illustrates the interplay among these categories, showing feedback loops where discrepancies trigger redesign or altered use. Examples include:

- If intended and embodied values align but realized values differ, changing use suffices.
- If embodied values diverge from intended ones, redesign is needed.
- Unintended consequences (e.g., an AI causing discrimination) may necessitate revising intended values to include fairness (p. 390).

Van de Poel emphasizes redesign as an ongoing process, especially for adaptive AI systems, involving not just designers but users and operators (p. 390).

AI Systems as Sociotechnical Systems: Five Building Blocks

Recognizing AI's complexity, van de Poel conceptualizes AI systems as sociotechnical systems, expanding beyond traditional models (Bauer & Herder, 2009; Baxter & Sommerville, 2011; Geels, 2004; Bruijn & Herder, 2009; Pasmore & Sherwood, 1978; Kroes et al., 2006; Ottens et al., 2006; Dam et al., 2013; Franssen, 2014; Nickel, 2013) (p. 391). Traditional systems comprise:

1. **Technical Artifacts**: Physical objects with designed functions (Kroes, 2010; Kroes & Meijers, 2006; Houkes et al., 2002; Houkes & Vermaas, 2010; Vermaas & Houkes, 2006) (p. 391).
2. **Human Agents**: Individuals with intentionality and moral agency (p. 391).
3. **Institutions**: Rules governing human behavior (North, 1990; Calvert, 1995; Ullmann-Margalit, 1977; Ostrom, 2005; Bicchieri, 2006) (p. 392).

AI systems add two unique components:

4. **Artificial Agents (AAs)**: Autonomous, interactive, and adaptive entities lacking human intentionality (Floridi & Sanders, 2004) (p. 392).
5. **Technical Norms**: Code-based rules regulating AAs, grounded in causal-physical terms (Mahmoud et al., 2014) (p. 392).

Table 1 (p. 391) distinguishes these blocks by intentional versus physical-causal natures, highlighting AI's hybridity and adaptivity as both opportunities and challenges for value embedding (Wallach & Allen, 2009; Anderson & Anderson, 2011; Cave et al., 2019; Vanderelst & Winfield, 2018) (p. 387).

Value Embedding in Technical Artifacts

Adapting his earlier work (Van de Poel & Kroes, 2014), van de Poel defines value embodiment in technical artifacts:

> "Technical artifact x embodies value V if the designed properties of x have the potential to achieve or contribute to V (under appropriate circumstances) due to the fact that x has been designed for V." (p. 393)

This hinges on two connected conditions: (1) design intent for V, and (2) conduciveness to V through use. He provides four examples:

1. **Sea Dikes**: "designed to protect against flooding... conducive for protection against flooding... therefore embody the value of safety" (p. 393).
2. **Bread Knife**: "designed to cut bread... can also be used for killing... does not embody the (dis)value of killing" because it lacks design intent (p. 393).
3. **Faulty Pacemaker**: "designed for (contributing to) human well-being... fails to contribute to well-being" due to poor design, thus not embodying well-being (p. 393).
4. **Recommender System**: "designed to serve its customers may unintendedly... contribute to filter bubbles and echo chambers... (dis)values such as lack of respect and untruth" are not embodied without redesign intent (p. 393-394).

For unintended consequences, he suggests redesign—by designers or users (Vermaas & Houkes, 2006)—can embed new values, distinguishing passive, idiosyncratic, and innovative use (p. 394). If negative outcomes persist, designers acquire an obligation to embed positive values (Winner, 1977) (p. 394).

Value Embedding in Institutions

Institutions, as rules, also embody values. Using the ADICO grammar (Crawford & Ostrom, 1995), van de Poel categorizes:

- **Shared Strategies (AIC)**: Descriptive expectations, e.g., "Pedestrians (A) use an umbrella to avoid getting wet (I) when it rains (C)" (p. 395).
- **Norms (ADIC)**: Deontic expectations without sanctions, e.g., "Residents (A) must (D) greet their neighbors (I) in this neighborhood (C)" (p. 395).
- **Rules (ADICO)**: Deontic expectations with sanctions, e.g., "Car drivers (A) must (D) drive on the right side of the road (I) in the Netherlands (C), otherwise they will be fined by the police (O)" (p. 395).

He proposes:

> "Institution R embodies value V if R is conducive to V because R has been designed for V." (p. 396)

Examples include:

1. **Traffic Rule**: Embodies safety as it's designed for and conducive to traffic safety (p. 396).
2. **Greeting Norm**: Embodies politeness through design and effect (p. 396).
3. **Pavement Strategy**: Embodies convenience via shared design and conduciveness (p. 396).

Like artifacts, institutions require both conditions, with redesign addressing unintended outcomes (p. 397).

Human Agents: Mediators of Value

Human agents—users, operators, designers—interact with embodied values in artifacts (V_T) and institutions (V_I), influenced by personal values (V_A) (p. 397). Figure 2 (p. 398) shows this intentional-causal dynamic, with outcomes potentially diverging from embodied values. Humans' reflective capacity (Figure 3, p. 399) enables evaluation and redesign, balancing system stability and adaptability (Franssen, 2015) (p. 397-398).

Artificial Agents: Designed Autonomy

AAs, distinct from humans and artifacts (Table 2, p. 400), embody values if designed for them and conducive to them (p. 399-400). Moor's (2006) taxonomy classifies AAs:

1. **Ethical Impact Agents**: Affect values without intent, thus not embodying them (p. 400).
2. **Implicit Ethical Agents**: Designed to follow values, embodying them if effective (p. 400).
3. **Explicit Ethical Agents**: Represent and reason about values, embodying them if top-down designed (p. 401).
4. **Full Ethical Agents**: Hypothetical with human-like traits, currently unfeasible (Winfield, 2019; Müller, 2020) (p. 400).

Adaptivity risks disembodying values, requiring restrictions or monitoring (Grodzinsky et al., 2008; Cervantes et al., 2020; Allen et al., 2005; Wallach & Allen, 2009; Cave et al., 2019; van Wynsberghe & Robbins, 2019) (p. 400-401).

Technical Norms: Regulating AAs

Technical norms, akin to institutions for humans, regulate AAs via code (Lessig, 1999; Akrich, 1992; Latour, 1992; Thaler & Sunstein, 2009; Fogg, 2003; Norman, 2000) (p. 401-402). Created offline or emergently (Hollander & Wu, 2011), they embody values if:

> "Technical norm N embodies value V if (1) N has been designed (by the human system designers) for V and (2) the execution of N within the system is conducive to V." (p. 402)

This mirrors the artifact and institution accounts (Mahmoud et al., 2014; Aldewereld & Sichman, 2013; Panagiotidi et al., 2013; Dybalova et al., 2014; Leenes & Lucivero, 2014) (p. 402).

System-Level Value Embedding

At the system level, van de Poel proposes:

> "Value V is embodied in sociotechnical system S if S is conducive to V because of those components of S that have been designed for V." (p. 403)

This accommodates systems not wholly designed (Bowker et al., 2010), requiring only some components to embody V, with conduciveness tied to following norms and use plans (p. 403-404).

Conclusion: Practical Lessons

Van de Poel concludes with two lessons:

1. **Continuous Monitoring and Redesign**: AI's adaptivity necessitates oversight and human control (Santoni de Sio & van den Hoven, 2018) to manage disembodiment risks (p. 404-405).
2. **Focus on Technical Norms**: Embedding values in norms may be more effective than in AAs alone, shifting focus from machine ethics to system regulation (p. 405).

Building Ethics into Artificial Intelligence by Han Yu et al.,

---

This analysis delves deeply into the content, structure, and contributions of the paper, incorporating all quotes, references, and examples provided in the document. It avoids being a mere summary by exploring the nuances, implications, and technical details of each section, while maintaining a comprehensive and self-contained narrative. Let's dive in.

Introduction: Setting the Stage for Ethical AI

The paper *"Building Ethics into Artificial Intelligence"* by Han Yu and co-authors from institutions like Nanyang Technological University, University of Massachusetts Amherst, and Hong Kong University of Science and Technology, tackles a pressing issue in modern AI research: how to ensure AI systems make ethical decisions. Published with a focus on technical solutions, it bridges a gap left by previous surveys that emphasized psychological, social, and legal perspectives over actionable computational approaches.

The abstract outlines the paper's intent clearly:

> "As artificial intelligence (AI) systems become increasingly ubiquitous, the topic of AI governance for ethical decision-making by AI has captured public imagination. Within the AI research community, this topic remains less familiar to many researchers. In this paper, we complement existing surveys... with an analysis of recent advances in technical solutions for AI governance."

This sets the stage for a technical exploration, distinct from the broader public discourse often dominated by fears of artificial general intelligence (AGI). The authors acknowledge this public anxiety:

> "A major source of public anxiety about AI, which tends to be overreactions [Bryson and Kime, 2011], is related to artificial general intelligence (AGI) [Goertzel and Pennachin, 2007] research aiming to develop AI with capabilities matching and eventually exceeding those of humans."

However, they quickly pivot to a more immediate concern: "Although we are still decades away from AGI, existing autonomous systems (such as autonomous vehicles) already warrant the AI research community to take a serious look into incorporating ethical considerations into such systems." This pragmatic focus on current systems, like autonomous vehicles (AVs), grounds the paper in real-world relevance.

The authors define ethics via Cointe et al. (2016), framing it as "a normative practical philosophical discipline of how one should act towards others," encompassing three dimensions:

1. **Consequentialist Ethics**: "An agent is ethical if and only if it weighs the consequences of each choice and chooses the option which has the most moral outcomes." Also called utilitarian ethics, it prioritizes aggregate benefits.
2. **Deontological Ethics**: "An agent is ethical if and only if it respects obligations, duties and rights related to given situations." This rule-based approach aligns with social norms.
3. **Virtue Ethics**: "An agent is ethical if and only if it acts and thinks according to some moral values (e.g. bravery, justice, etc.)." It emphasizes an intrinsic moral character.

They further define ethical dilemmas as "situations in which any available choice leads to infringing some accepted ethical principle and yet a decision has to be made [Kirkpatrick, 2015]." This foundational framework underpins the paper's taxonomy of AI governance techniques, divided into four areas:

- **Exploring Ethical Dilemmas**
- **Individual Ethical Decision Frameworks**
- **Collective Ethical Decision Frameworks**
- **Ethics in Human-AI Interactions**

Each area is explored in depth below, with every detail, quote, and reference meticulously unpacked.

Exploring Ethical Dilemmas: Understanding Human Preferences

The first area, "Exploring Ethical Dilemmas," focuses on tools that help the AI community understand human preferences in ethical scenarios. The paper states: "In order to build AI systems that behave ethically, the first step is to explore the ethical dilemmas in the target application scenarios." Two key tools are highlighted: GenEth and the Moral Machine project.

GenEth: Expert-Driven Ethical Analysis

Proposed by Anderson and Anderson (2014), GenEth is an "ethical dilemma analyzer" designed to involve ethicists in codifying ethical principles for AI. The authors note:

> "They realized that ethical issues related to intelligent systems are likely to exceed the grasp of the original system designers, and designed GenEth to include ethicists into the discussion process in order to codify ethical principles in given application domains."

GenEth employs a structured representation schema:

1. **Features**: "Denoting the presence or absence of factors (e.g., harm, benefit) with integer values."
2. **Duties**: "Denoting the responsibility of an agent to minimize/maximize a given feature."
3. **Actions**: "Denoting whether an action satisfies or violates certain duties as an integer tuple."
4. **Cases**: "Used to compare pairs of actions on their collective ethical impact."
5. **Principles**: "Denoting the ethical preference among different actions as a tuple of integer tuples."

This framework is operationalized through a graphical user interface, where discussions are processed using inductive logic programming to "infer principles of ethical actions." GenEth's strength lies in its systematic approach, leveraging expert input to formalize ethics, though it lacks the scalability of crowd-based methods.

Moral Machine: Crowdsourcing Ethical Preferences

In contrast, MIT's Moral Machine project (http://moralmachine.mit.edu/) uses crowdsourcing to gather public opinions on ethical dilemmas, particularly for AVs. The paper explains:

> "The Moral Machine project focuses on studying the perception of autonomous vehicles (AVs) which are controlled by AI and has the potential to harm pedestrians and/or passengers if they malfunction."

Participants judge scenarios, such as whether an AV should sacrifice its passenger to save more pedestrians, with preferences analyzed across eight considerations:

1. Saving more lives
2. Protecting passengers
3. Upholding the law
4. Avoiding intervention
5. Gender preference
6. Species preference
7. Age preference
8. Social value preference

The project's findings, based on 3 million participants, reveal a nuanced public stance:

> "Based on feedbacks from 3 million participants, the Moral Machine project found that people generally prefer the AV to make sacrifices if more lives can be saved. If an AV can save more pedestrian lives by killing its passenger, more people prefer others' AVs to have this feature rather than their own AVs [Bonnefon et al., 2016; Sharif et al., 2017]."

This highlights a consequentialist bent—favoring the greater good—but also a self-interest paradox, as individuals hesitate to apply the same logic to their own vehicles. The authors caution:

> "Nevertheless, self-reported preferences often do not align well with actual behaviours [Zell and Krizan, 2014]. Thus, how much the findings reflect actual choices is still an open question."

Alternative suggestions emerge, such as random decision-making ("let fate decide") per Broome (1984) or segregating AVs from human traffic (Bonnefon et al., 2016). These diverse opinions underscore the complexity of ethical consensus, making tools like Moral Machine critical yet imperfect for informing AI design.

Individual Ethical Decision Frameworks: Empowering Single Agents

The second area, "Individual Ethical Decision Frameworks," explores mechanisms for single AI agents to make ethical decisions. The paper asserts: "When it comes to ethical decision-making in AI systems, the AI research community largely agrees that generalized frameworks are preferred over ad-hoc rules." Several frameworks are detailed, each with unique approaches.

MoralDM: Combining Rules and Analogies

Dehghani et al. (2008) propose MoralDM, which integrates:

1. **First-Principles Reasoning**: "Makes decisions based on well-established ethical rules (e.g., protected values)," such as the moral unacceptability of murder regardless of outcome.
2. **Analogical Reasoning**: "Compares a given scenario to past resolved similar cases to aid decision-making."

The authors note: "As the number of resolved cases increases, the exhaustive comparison approach by MoralDM is expected to become computationally intractable." Thus, Blass and Forbus (2015) extend it with "structure mapping," which "trims the search space by computing the correspondences, candidate inferences and similarity scores between cases." This enhancement improves efficiency while retaining MoralDM's ability to handle culturally sensitive "protected values."

BDI-Based Ethical Judgment

Cointe et al. (2016) offer a framework using the Belief-Desire-Intention (BDI) model (Rao and Georgeff, 1995), structured around:

- **Awareness**: "Generates the beliefs that describe the current situation facing the agent and the goals of the agent."
- **Evaluation**: "Generates the set of possible actions and desirable actions."
- **Goodness**: "Computes the set of ethical actions based on the agent's beliefs, desires, actions, and moral value rules."
- **Rightness**: "Evaluates whether or not executing a possible action is right under the current situation."

This process adapts to judging others' actions under varying information levels (blind, partially informed, fully informed), though it lacks "quantitative measure of how far a behaviour is from rightfulness or goodness," limiting its precision.

Game Theory and Machine Learning

Conitzer et al. (2017) propose two approaches:

1. **Game Theory**: Extends the "extensive form" (game trees) with "passive actions" to account for protected values, addressing scenarios where inaction is ethical.
2. **Machine Learning**: Classifies actions as right or wrong using human judgments, potentially from Moral Machine data, though cultural inconsistencies pose challenges. They suggest leveraging moral foundations (e.g., harm/care, fairness) from Clifford et al. (2015) for generalizable representations.

The authors envision combining these: "Game theory and machine learning can be combined into one framework in which game theoretic analysis of ethics is used as a feature to train machine learning approaches."

CP-Nets for Preference Balancing

Loreggia et al. (2018) use CP-nets to "represent the exogenous ethics priorities and endogenous subjective preferences," introducing a "notion of distance between CP-nets" to balance agent preferences with ethical requirements. This allows flexibility in decision-making when preferences align closely with ethics.

High-Level Action Language

Berreby et al. (2017) shift moral reasoning to agents via a high-level action language, implemented with answer set programming (Lifschitz, 2008). It:

- Simulates outcomes using action, event, and situation data.
- Produces causal traces via a causal engine.
- Assesses goodness and rightfulness using ethical specifications and deontological rules.

This framework enables agents to "decide and explain their actions, and reason about other agents' actions on ethical grounds," reducing the burden on developers.

Ethics Shaping in Reinforcement Learning

Wu and Lin (2018) adapt reinforcement learning (RL) with "ethics shaping," incorporating ethical values into reward functions:

> "By assuming that the majority of observed human behaviours are ethical, the proposed approach learns ethical shaping policies from available human behaviour data in given application domains."

The function "rewards positive ethical decisions, punishes negative ethical decisions, and remains neutral when ethical considerations are not involved," separating ethics from standard RL design.

Collective Ethical Decision Frameworks: Group Dynamics

The third area, "Collective Ethical Decision Frameworks," addresses ethical decision-making among multiple agents. Pagallo (2016) argues: "Individual ethical behavior isn't enough and that we need social norms and rules that can evolve."

Social Norms and Trust Networks

Singh (2014; 2015) proposes a distributed framework using social norms, defined via:

- Roles (qualifications, privileges, penalties)
- Commitments, authorizations, prohibitions, sanctions, and power

Agents form "a network of trust based on techniques from the reputation modelling literature [Yu et al., 2010; Yu et al., 2013]" for self-governance.

Human-Agent Collectives

Greene et al. (2016) envision agents evaluating different ethical dimensions (deontological, consequentialist, virtue) within human-agent collectives (Jennings et al., 2014). Preferences are aggregated, but challenges include:

- Large action sets outnumbering agents
- Interdependent actions
- Missing or imprecise preferences

Voting-Based System

Noothigattu et al. (2018) build on Moral Machine data, using a voting system with "swap-dominance" to rank alternatives:

> "Assuming everything else is fixed, an outcome a swap-dominates another outcome b if every ranking which ranks a higher than b has a weight which is equal to or larger than rankings that rank b higher than a."

This ensures computationally efficient, consequentialist decisions reflecting collective preferences.

Ethics in Human-AI Interactions: Influencing Behavior

The fourth area, "Ethics in Human-AI Interactions," focuses on AI influencing humans ethically, guided by the Belmont Report (Bel, 1978):

1. "People's personal autonomy should not be violated."
2. "Benefits brought about by the technology should outweigh risks."
3. "The benefits and risks should be distributed fairly among the users."

Persuasion Agents

Stock et al. (2016) study AI persuasion in the trolley problem, testing:

1. Emotional appeals
2. Utilitarian arguments

3. Lying

Findings show: "Participants hold a strong preconceived negative attitude towards the persuasion agent, and argumentation-based and lying-based persuasion strategies work better than emotional persuasion strategies."

Emotional Responses

Battaglino and Damiano (2015) use Coping Theory (Marsella and Gratch, 2003) to trigger emotions like shame (for self-violations) or reproach (for others' violations), enhancing human-AI interaction.

Discussions and Future Directions

The paper notes a focus on individual frameworks, a need for diverse cultural data, and challenges in collective preference representation and human-AI ethics. It advocates interdisciplinary collaboration and a global AI regulatory framework (Erdélyi and Goldsmith, 2018). Future directions include:

1. **Social-Systems Analysis**: Using transfer learning (Pan and Yang, 2010) to model diverse ethics.
2. **Revising Social Contracts**: Dynamic regulations for AI responsibility.
3. **Explainable AI**: Argumentation-based explanations (Fan and Toni, 2015) balancing transparency.
4. **Adversarial Considerations**: Incorporating adversarial game theory (Vorobeychik et al., 2012) to counter strategic exploitation.

This analysis covers every facet of the paper, from its foundational definitions to its technical proposals, ensuring a thorough understanding of how ethics can be built into AI.