Wasabi: A Conceptual Model for Trustworthy AI

Amika M. Singh
Harvard Law School

Munindar P. Singh
NC State University

***Abstract*—The expansion of AI into our lives and livelihoods makes clear that we must develop AI to be ethical and trustworthy. We propose Wasabi, a novel conceptual model for trustworthy AI based on an adaptation of the well-known ability-benevolence-integrity model of trust to trustworthiness.**

Current approaches to trustworthy AI propose lists of desirable properties, including fairness, explainability, and accountability. However, these properties inadequately cover the criteria of ability, benevolence, and integrity, and the resulting incompleteness hurts trustworthiness even when these properties are met.

We examine case law as evidence for concepts underlying trustworthiness. Legal cases represent boundary conditions that were vigorously contested by lawyers and carefully deliberated on by juries. Thus, they capture important details and tradeoffs absent in shallower analyses. From each case, we identify lessons for AI. We close with directions for future investigation.

Index Terms: Responsible AI, Trust in AI, Trustworthy technology

1. Introduction

Trust is crucial to any interaction between two or more autonomous entities because their very autonomy leaves each entity potentially vulnerable to the decisions of the others. With the expansion of AI, it is crucial that the AI is trustworthy, not merely trusted by users.

Trust, broadly, is relational: the trusting party or *trustor* willingly makes itself vulnerable to the trusted party or *trustee* [1]. We take the scope of trust to include settings where the trustor may practically lack a choice about whether to deal with the trustee. These settings are common in AI applications. For example, prospective borrowers cannot choose what AI algorithms assess their loan applications. Sometimes, the stakeholders collectively may have a say, e.g., through regulatory mechanisms, but such control is not guaranteed.

We understand trustworthiness as the correlate or the "flip-side" of trust. Specifically, it is not concerned with the beliefs of the trustor about the trustee but with the relevant attributes of the trustee. Trustworthiness can be approached from the conception of reliability, but we posit it's better characterized in the same way as trust. To this end, we adapt Mayer et al.'s [2] framework for trust by turning it around to focus on trustworthiness.

How trust and trustworthiness are conceived, especially in connection with AI, involves multiple key dimensions. One dimension is whether the AI is embodied or not, and to the extent it is or pretends to be human-like. That is, the AI's apparent independence in functioning would raise expectations of how trustworthy it needs to be. For example, though created, maintained, and managed by humans, an autonomous vehicle drives by itself and thus differs from a mortgage loan approval program, which merely helps a bank officer.

A second dimension is who the trustor and trustee are. For example, for sentencing software, the trustors include convicts, judges, and society at large, and the trustees include the AI, its developer, and the judicial system. For autonomous vehicles, the trustors include passengers, pedestrians, occupants of other vehicles, and the transportation authority, and the trustees include the vehicle, developers, operators, and relevant safety boards. This relational conception of trustworthiness turns out to be a major extension beyond current approaches.

Third, trust can be directed from a trustor to a single entity (e.g., an AI) or to the sociotechnical system (STS) [3] in which both trustor and trustee function. An STS is characterized by the norms—informal social norms as well as laws and regulations—between its members and provides the social context in which its members interact. An STS can be viewed as an entity in its own right, separate from its members [4]. Through its norms, an STS induces trustworthiness in its members and potentially advances compensation for those harmed. Thus, an STS can be trustworthy even if some of its members are not.

Contributions

Against this backdrop, we make three contributions by

- proposing the Wasabi model of trustworthiness centered on a trustee promoting the goals, interests, and values of a trustor;
- validating the Wasabi model through an analysis of legal precedent; and
- identifying gaps in popular models of trustworthy AI.

2. The Wasabi Model of Trustworthiness

Mayer et al. [2] proposed one of the leading models of trust based on three core concepts. *Ability*, *Benevolence*, and *Integrity* respectively refer to the trustor's belief in the trustee's capabilities to perform a task as desired, in the trustee's intention to help the trustor, and in the trustee's moral character. This "ABI" model has been validated in numerous studies.

Taking inspiration from it, we propose the Wasabi model, which flips the ABI model around and focuses on the trustee instead of the trustor. Whereas the ABI model focuses on the subjective state of the trustor and has nothing to do with the trustee's true nature, Wasabi characterizes the trustee itself. That is, Wasabi incorporates the beliefs, goals, and values of the trustee with respect to the trustor though it elides the trustor's beliefs about those beliefs, goals, and values.

In current thinking, the trustworthiness of AI is a property of an individual (trustee). In contrast, in the Wasabi conception, trustworthiness is relational, i.e., a property of a trustee with respect to a trustor given a context and purpose. Specifically, trustworthiness ratings would fall on a spectrum, and where they fall would depend on the context and purpose.

The Wasabi dimensions aren't perfectly orthogonal but capture broadly distinct intuitions: Can the trustee do its job? Does the trustee seek to look after the trustor's interests? Will the trustee respect relevant societal norms and values? In conceptual terms, these questions ask whether the trustee will promote the trustor's goals, interests, and values [5].

Figure 1 illustrates the Wasabi model, highlighting these three components along with some example properties the trustee ought to possess to be trustworthy; ideally, the trustee would authentically project those properties and gain trust accordingly from the trustor.

Trust is sometimes applied to technical entities, such as machines. However, researchers have shown that "trust" in a machine viewed as an instrument is a weaker construct than trust in a sociocognitive entity, such as a person [1]—e.g., blame applies to a person but not to a machine. Accordingly, we consider AI not as a standalone technical artifact but as part of an STS [3].

The Wasabi model applies to STSs, encompassing both AI and humans and a society. Even though people may treat AI differently from humans, to be trustworthy, an AI in an STS should be held to the same norms and expectations as humans.

3. Case Law as Empirical Evidence

Trust and trustworthiness are subtle concepts. Their definitions cannot be imposed by technologists and instead must be established based on empirical evidence of how the public *and* experts understand them. But their subtlety makes such empirical investigation difficult.

We position case law as an empirical source of the public's understanding. For jury trials, the jury is selected from the public and, for important cases, would deliberate on the order of days, knowing that money and personal freedom are at stake. Litigation would bring out the relevant facts and context and challenge them through expert and other testimony. For appellate cases, where there is no jury, the judges involved would conduct an in-depth review of the facts and relevant doctrines. Thus, for a complex situation, we can place greater confidence in the outcome of a legal case than we might in surveys of the public based on hypothetical situations.

Moreover, because only extreme cases make it to litigation, this approach coheres with the well-known *critical incident technique* [6], invented in the US military, which involves analyzing extreme cases as a way to derive criteria for "typical" performance.

Interestingly, whereas the law has little to say about whom someone should trust, it does have a lot to say about who is trustworthy. Approaching trustworthiness from the standpoint of the law provides clearer intuitions regarding (1) how the Wasabi components relate to interactions involving people and organizations and (2) principles to ground the understanding of trustworthy AI.

4. The Wasabi Model and Case Law

It's still early days for AI, and whereas there are public outcries and some litigation about excessive data collection and manipulation in social media, case law in AI is recent and not yet settled. We return to this point in the Conclusions.

However, since AI has some human-like qualities and seeks to help people and organizations, we find that existing case law does cover the challenges of trustworthiness that Wasabi brings forth. Below, we discuss the relevant case law, roughly in order of increasing complexity.

4.1. Ability: Unintended Side Effects

Just like for animals, the owners or operators of AI agents may not be able to predict or control their behavior, e.g., because of the complexity of algorithms or dependence on training data: even so, the owners remain liable for their actions.

In Baker v. Howard County Hunt (https://www.courtlistener.com/opinion/3486934/baker-v-howard-county-hunt/), hounds running with the members of a hunting club entered Baker's property while pursuing a fox and proceeded to kill some of Baker's animals.

The court held that the hunters were liable for the actions of the hounds, especially as this was not an isolated incident, and they apparently knew of the hounds' behavior. By hunting near Baker's property, the hunters risked trespass by the hounds and the subsequent loss to Baker even though they didn't instruct the hounds to attack Baker's livestock and assumed the hounds had "fidelity" to the chase.

Insight: When one's agents pose a risk to someone out of the ordinary, there's a corresponding liability.

AI Lesson 1. Ability: Risky actions

Trustworthy AI must avoid imposing externalities of an action on others.

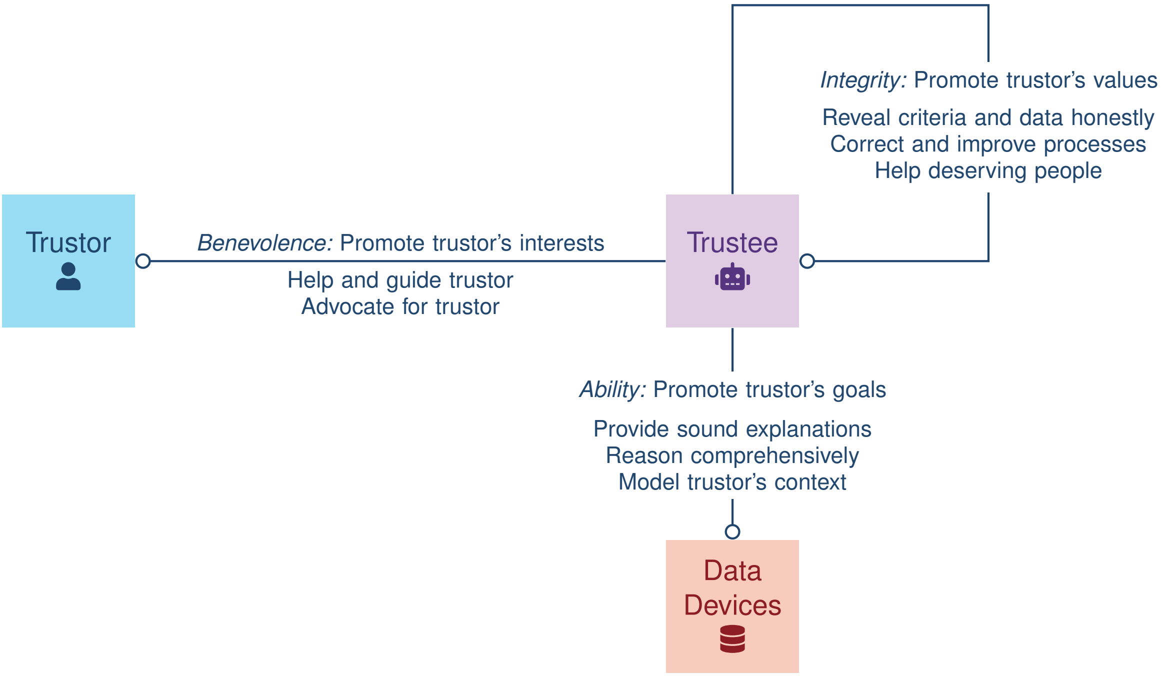
Trustor
Benevolence: Promote trustor's interests
Help and guide trustor
Advocate for trustor
Trustee
Integrity: Promote trustor's values
Reveal criteria and data honestly
Correct and improve processes
Help deserving people
Ability: Promote trustor's goals
Provide sound explanations
Reason comprehensively
Model trustor's context
Data
Devices

Figure 1. The Wasabi model: Capturing trustworthiness in AI in terms of ability, benevolence, and integrity.

4.2. Ability: Design Process

Grimshaw, a landmark case in US tort law, is demonstrative of how ability may affect liability. (https://law.justia.com/cases/california/court-of-appeal/3d/119/757.html). In 1970, Ford released a new compact model, the Pinto. In 1972, Lilly Gray was driving a Ford Pinto when she was rear-ended by another vehicle. The collision caused the gas tank to rupture, releasing gas into other chambers of the car, leading the car to explode in fire. Gray died from congestive heart failure resulting from her burns; her passenger, Grimshaw, survived after undergoing many surgeries for his burns, though he suffered the loss of fingers and an ear.

Grimshaw and Gray's families sued Ford for damages caused by the design defect that caused the engine to rupture on impact. The court considered the industry standard of how to should respond to safety failures to determine Ford's liability: they determined that the standard of care after a failed safety test was to "redesign and retest," and Ford, knowing of the design defect in the Pinto, went ahead with selling and marketing the car without taking those requisite steps.

Insight: It's not just the outcome (e.g., an accident) that determines liability, but the departure from established practice regarding the resources to allocate toward preventing or repairing the defects that cause the outcome. (We return to this case in the integrity discussion below.)

AI Lesson 2. Ability: Designing against risks

Trustworthy AI requires design processes that ensure robustness against threats in the operating environment, including combinations of rare events that prove hazardous.

4.3. Benevolence: Unintended Effects of Malfeasance

A major concept in tort theory is intent, which is relevant to the *benevolence* factor of the ABI model. An intentional tort is done with intent on the part of the *tortfeasor* (the wrong-doer). Intent does not necessarily mean they intended to create harm, simply to complete the action that led to harm.

If the action that caused harm was intended, even if the resulting harm was not, the tortfeasor is liable. Vosburg v. Putney (https://h2o.law.harvard.edu/cases/2451) is a quintessential illustration from 1890. Putney, 11 years old, kicked Vosburg, then 14 years old, in the shin. Putney did not know that Vosburg had a previous injury and that the kick would cause him to develop a serious infection and limit the use of his leg for the rest of his life. The Supreme Court of Wisconsin held Putney liable: he did not intend to cause Vosburg's severe injury but nonetheless intended to commit battery by kicking Vosburg. Although Vosburg's context (his health) was unknown to Putney, that didn't relieve Putney of his liability.

Insight: An actor is accountable even for consequences of their actions they did not foresee or lacked the information to foresee.

AI Lesson 3. Benevolence: Accountability in context

Trustworthy AI would learn about and accommodate the trustor's context and ensure its decision is in their interest in that context.

4.4. Benevolence: Duty of Care

The duty of care is a prominent consideration in corporate law. The law creates a duty of care on directors because they make important decisions that have a material impact on the shareholders. The shareholders largely have no option but to accept their decisions. Moreover, directors and shareholders have shared interests in maintaining or increasing the valuation of a company.

In Graham v. Allis-Chalmers Manufacturing Company (https://law.justia.com/cases/delaware/supreme-court/1963/188-a-2d-125-3.html), Allis-Chalmers and four of its managers (not directors), pleaded guilty to price-fixing in violation of federal antitrust laws. The board of directors met annually but did not discuss pricing. The shareholders suffered when the company was held liable for its price fixing. Two shareholders filed for a derivative action against the company's directors to recover what they lost from the price-fixing litigation. Although the directors did not have knowledge of the illegal acts, the Delaware Supreme Court considered whether their lack of knowledge meant they had breached their duty of care. It held that the directors did not breach their duty of care because they could not be expected to suspect foul play without compelling reason to do so. That is, corporate law provides weak mechanisms to enforce a trusting relationship, falling short of enforcing trustworthiness.

Insight: A duty of care may be insufficient for a finding of negligence in anomalous cases where the party may not be expected to obtain the requisite knowledge to identify violations.

AI Lesson 4. Benevolence: Obtaining requisite knowledge

Trustworthiness requires not only supporting a trustor's interests, but also acquiring the knowledge to determine if those interests are at risk.

The duty of care in tort law offers a clearer framework for autonomous agents than corporate law. Importantly, a duty of care may arise on the fly when an actor engages in an action that may cause a risk to others. For example, whereas an autonomous vehicle has a special relationship with a passenger, it also has a duty of care to a pedestrian crossing in front—someone who just happens to be there.

For example, in United States v. Lawter (https://casetext.com/case/united-states-v-lawter), the court found a duty to not injure someone while helping them, even if helping them was a gratuitous act. Specifically, the Coast Guard attempted to rescue Loretta Lawter when her boat sank. They lowered a cable from a helicopter and attempted to pull her up. However, they failed to follow the standard procedure and didn't secure her to the cable, leaving her to simply hold on to it. She was unable to do so and fell to her death. The court found that by attempting to save her, the Coast Guard had assumed a duty of care, which it failed to meet.

Insight: Failing to rise to the duty of care required by the situation at hand constitutes negligence.

AI Lesson 5. Benevolence: Care in context

Trustworthiness presumes identifying and honoring duties of care that arise on the fly and viewing incidental stakeholders as trustors.

4.5. Conflicts in Trustworthiness

There is a general expectation that agents will serve clients' interests. In most scenarios, as the US Supreme Court noted in Jaffee v. Redmond (https://caselaw.findlaw.com/us-supreme-court/432/43.html), information collected during an individual's therapy session is protected, and the therapist cannot be legally compelled to give that information to the Court. In this case, the notes may have implicated Redmond, a police officer who had shot and killed someone. A jury awarded damages in the initial outcome (without having seen the notes). That decision was thrown out because the judge had indicated that the notes should have been provided and that the therapist was wrong to withhold them. A therapist who is trustworthy to their client maintains confidentiality of the client's information. The duty here is on the part of the therapist to their patient.

Tarasoff v. Regents of University of California (https://law.justia.com/cases/california/supreme-court/3d/17/425.html) shows that there might be a duty greater than that between the therapist and their client. In Tarasoff, the plaintiffs were the parents of the deceased Tatiana Tarasoff, whom her classmate, Poddar, had killed. Two months before her murder, Poddar confided his intent to kill Tarasoff to his psychiatrist. The Supreme Court of California held that the plaintiffs could state a claim against the psychiatrist because the psychiatrist did not warn Tarasoff or her parents or do anything to confine Poddar to prevent him from murdering her. The psychiatrist was negligent in failing to warn. In this case, the psychiatrist failed to be trustworthy to society.

Clearly, to be trustworthy to one's client conflicts here with being trustworthy to society. The greater duty here was to prevent an innocent person from being killed. Autonomous agents ought to be able to work under a similar trustworthiness framework, balancing competing duties and selecting the right one to uphold, which might involve comparing their responsibilities and potential blameworthiness [7].

Insight: Competing duties—such as protection of one's client and disclosure of risk to others—must be compared concerning the severity of potential harm.

AI Lesson 6. Trustors can conflict

An effective AI governance framework would address conflicts between the interests of the relevant trustors, potentially based on societal values.

4.6. Integrity and Transparency

Whereas the product defect in the Pinto corresponds to a lack of ability, Ford's apparent lack of disclosure about the defect in the gas tank placement corresponds to a lack of integrity. Whereas the ability violation led to compensatory damages for product defects, the integrity violation led to additional punitive damages.

Schwartz's [8] analysis of Grimshaw explains how "purchasers generally lack knowledge of specific hazards that inhere in the products' designs" and "the confidentiality of Ford's life-affecting design choices is an important part of the ethical dimension of the Pinto case myth" (p. 1068). Ford's culpability was based not on the Pinto's faulty design alone but on whether Ford knew that such an impact would lead to a punctured gas tank.

Note that Schwartz describes the failure of Ford—that is, he focuses on the trustworthiness of the *individual* (here, Ford). In contrast, Gladwell [9] describes US Congressional hearings led by Tim Murphy, Chairman of the House Subcommittee on Oversight and Investigations. Murphy's focus was "not just to fix the car but to fix a culture within a business and a government regulator that led to these problems" and "the fact that the National Highway Traffic Safety Administration did not order it recalled in a timely fashion represented a moral failure." This is a clear case of the lack of *systemic* trustworthiness.

Insight: Regulations can foster systemic trustworthiness through effective governance by requiring transparency between the key stakeholders.

AI Lesson 7. Integrity and openness

Trustworthiness is enhanced by regulations against withholding facts material to governance and stakeholder decision making.

4.7. Benevolence and Integrity: Acting honorably

Corporate law finds a high expectation of benevolence and integrity in partnerships. In Meinhard v. Salmon [10], Meinhard and Salmon had a joint venture where they leased the Bristol Hotel for 20 years. This partnership constituted a principal-agent relationship as Meinhard put up most of the capital for the venture while Salmon managed the business. The partnership was set to end with the end of the lease. As the lease was about to end, the owner of the property approached Salmon to negotiate a new deal, not realizing that Salmon was in a joint venture with Meinhard. Salmon signed the deal without Meinhard's participation. Meinhard sued, arguing that the new deal belonged to the joint venture. The court agreed with Meinhard, giving him 49 percent of the new venture. That is, the joint venture was such a strong bond between the two partners that Salmon broke his fiduciary duty by creating a new deal without informing Meinhard of it.

This case indicates how integrity and benevolence interrelate. Judge Benjamin Cardozo's legendary phrasing—"A trustee is held to something stricter than the morals of the market place. Not honesty alone, but the punctilio of an honor the most sensitive, is then the standard of behavior"—makes this point [10].

Insight: Duties to another party, especially in special relationships, yield higher standards for benevolence and integrity than the default.

AI Lesson 8. Depth of experience

The longer and deeper an AI interacts with a stakeholder, the higher the demand to advance the trustor's interests and respecting societal values.

4.8. Integrity: Governance

Conflicts of interest can subvert the integrity of a decision process, and the law takes such conflicts seriously.

ASME v. Hydrolevel (https://supreme.justia.com/cases/federal/us/456/556/) concerns members of the American Society of Mechanical Engineers, a nonprofit organization that creates safety standards for industrial products. Two "highly placed volunteers" of ASME (including a subcommittee chairman) were employed by a company whose products were in ASME's domain. These members manipulated the standards to render their competitor Hydrolevel's new product noncompliant.

The court decided that ASME was liable for violating federal antitrust laws because the ASME members were acting under its apparent authority when they colluded to alter the standards. Importantly, these members were not authorized by ASME to do what they did. Yet, they used their apparent authority under ASME—being trusted by ASME and, through ASME, by the industry at large—to make self-interested choices that violated antitrust laws.

Once they were caught, these members and their employer had already settled with Hydrolevel, acknowledging the clear lack of benevolence exhibited by them.

Therefore, this legal case didn't hinge on benevolence but on ASME's culpability arising from its poor governance that yielded a lack of integrity in its decision processes. The treble damages assessed indicate the moral outrage of the judge.

Insight: Inadequate governance to control members' behaviors and guide system outcomes constitutes a lack of integrity.

AI Lesson 9. Integrity of decision process

A governance framework must ensure that the AI (and relevant human administrators or operators) don't manipulate the decision process to serve their own interests over those of the trustor and that applicable laws and regulations flow down to the AI.

4.9. Integrity: Power Disparity

The law often finds that a duty to respect others' civil rights overrides personal interests. Housing law exemplifies this, with many statutes and cases supporting the notion that discrimination by a private individual or company is illegal. In Alexander v. Riga (https://caselaw.findlaw.com/us-3rd-circuit/1074323.html), the court found that a Black couple had been discriminated against when the owners of an apartment building lied to them because of their race, saying that no units were available.

By operating a rental business, the Rigas had assumed a position of power over tenants and were held to meeting societal values. The court held that Riga would have to pay compensatory and punitive damages, the latter being a clear indicator of an integrity violation.

Insight: Even private actors have the duty to respect the rights of others.

> **AI Lesson 10. Integrity: power disparity**
>
> Where the AI has greater power, to be trustworthy, it must meet a correspondingly higher bar in respecting fundamental societal values.

5. Extant Conception of Trustworthy AI

Current thinking formulates the trustworthiness of AI via a list of properties in the expectation that an AI that satisfies those properties is trustworthy. For example, Deloitte [11] lists fair and impartial, transparent and explainable, responsible and accountable, robust and reliable, respectful of privacy, and safe and secure. IBM [12] offers a similar list: explainability, fairness, robustness, transparency, and privacy.

The Wasabi model seeks not to replace such lists but to offer a more complete picture. The above properties are indeed all desirable, but they do not cover the gamut of the Wasabi model. Their general emphasis is on ability (promoting the trustor's goals, such as robustness and understanding of operations) and some components of integrity (promoting values such as fairness). Specifically, benevolence remains a challenge, because stakeholders may have conflicting interests.

We analyzed current conceptual definitions [11, 12] of the above-mentioned properties with respect to the Wasabi criteria. We identified how each property relates to Figure 1. Figure 2 visualizes our intuitions by allocating the contributions of each property to Wasabi's three components.

We briefly motivate these intuitions as follows. Robustness addresses reliable performance in the face of adversaries; it concerns how the AI deals with data and devices. Transparency means exposing algorithms and data usage to users; it emphasizes ability and has small benevolence and integrity components. Explainability seeks to help users understand AI decisions; explainability earns a higher benevolence component since it potentially helps a trustor improve their outcomes. Accountability concerns organizational and policy means to hold an entity to account [13], though it is frequently confused with traceability, e.g., [11]; accountability incorporates high benevolence and integrity since it helps trustors ensure the AI is well-behaved and helps advances organizational integrity through continual improvement. Privacy concerns safeguarding data through the lifecycle; in this limited sense, privacy focuses on devices and data but offers a component of benevolence. Fairness is about equity and mitigating bias; it focuses on ability in data processing with a significant integrity component by promoting the above values.

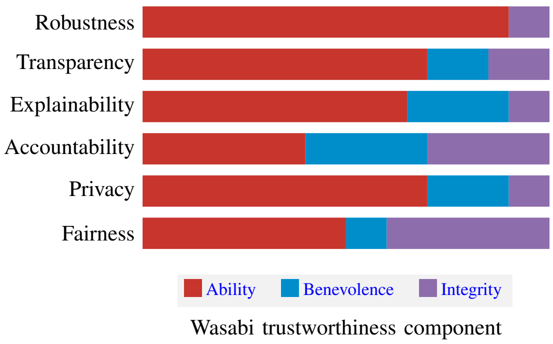

Figure 2. Visualizing the authors' intuitions of how traditional trustworthy AI properties associate with ability, benevolence, and integrity.

We make three important observations about current trustworthy AI approaches. First, they tend to be heavy on technology, with protecting user interests or societal values receiving less attention than they deserve. Second, there are many moral values [5], yet these approaches emphasize a few (e.g., fairness and accountability); a general framework would not arbitrarily restrict the values that stakeholders may hold dear. Third, sometimes when companies talk of risk, the risk they mean is to the provider's reputation [11, para 3], not primarily the risks to users or the public. This may explain their emphases on aspects of trustworthiness most likely to place the provider at risk.

6. Conclusions and Directions

The Wasabi model obtains support from legal theory and case law. Whereas the greatest emphasis in the law is on integrity and the least on ability, current conceptions of trustworthy AI are weighted heaviest on ability. Whereas ability may be more straightforward than benevolence and integrity, the contrast in attention is notable.

Bryan et al. [14] highlight important recent lawsuits pertaining to data collection. Although case law is newly emerging, and these cases largely concern the behavior of human actors in data collection, applying the Wasabi model helps understand their ramifications on trustworthiness. Most of these cases turn upon a lack of informed consent by the people whose data was collected, indicating a failure of integrity. None of these cases involved ability, which would be a case of malfunctioning biometrics. One case, involving a railroad scanning a truck driver's biometrics without consent (integrity violation), indicates a conflict with the public's interest in (and federal statutes for) railroad safety and security (indicating benevolence toward the public, though not toward the employee).

Trustworthiness goes beyond what is legally required. Case law addresses extreme situations, but in the spirit of critical incident analysis [6], we derive lessons for typical situations from these extreme situations. Important lessons for trustworthy AI include designing AI to anticipate and resist potential risks, understanding and accommodating each trustor's specific context, accommodating incidental stakeholders, modeling societal values to do the right thing when trustors' interests conflict, and avoiding misleading trustors. Whereas Wasabi doesn't provide predetermined answers about conflicts, it exposes the need for the AI to model the various stakeholders' goals, interests, and values and for methodologies to identify potential resolutions by bringing forth such conflicts to the stakeholders during design.

Additionally, important "meta" effects emerge: The potential appearance or portrayal of authority requires greater diligence in ensuring the decision process is not corrupted. The depth of engagement with a trustor and power over a trustor raise the standard for trustworthiness.

The domains of tort and corporate law are relevant to trustworthiness and AI practitioners would do well to gain awareness of them. So we hope the above illustrations prove useful besides lending support to the Wasabi model.

Our research opens some interesting directions for future research, including relating trustworthy AI to reasoning about value alignment [15, 16], design of assisting agents [17], negotiation of sociotechnical systems [18], and certification regimes [19]. Certification is needed to avoid an over-reliance, as in tort law, on "moments of carelessness" that randomly expose flaws in some agents [20].

Acknowledgments

MPS thanks the NSF (grant IIS-2116751) for partial support for this research.

REFERENCES

1. C. Castelfranchi and R. Falcone, *Trust Theory*. Chichester, UK: Wiley, 2010.
2. R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," *Academy of Management Review*, vol. 20, no. 3, pp. 709–734, 1995.
3. A. K. Chopra and M. P. Singh, "Sociotechnical systems and ethics in the large," in *Proc. AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*. New Orleans: ACM, 2018, pp. 48–53.
4. M. P. Singh, "Group ability and structure," in *Decentralized Artificial Intelligence, Volume 2*, Y. Demazeau and J.-P. Müller, Eds. Amsterdam: Elsevier/North-Holland, 1991, pp. 127–145.
5. S. H. Schwartz, "An overview of the Schwartz theory of basic values," *Online Readings in Psychology and Culture*, vol. 2, no. 1, pp. 3–20, 2012.
6. J. C. Flanagan, "The critical incident technique," *Psychological Bulletin*, vol. 51, no. 4, pp. 327–358, 1954.
7. V. Yazdanpanah, E. H. Gerding, S. Stein, C. Cîrstea, m. c. schraefel, T. J. Norman, and N. R. Jennings, "Different forms of responsibility in multiagent systems: Sociotechnical characteristics and requirements," *IEEE Internet Computing*, vol. 25, no. 6, pp. 15–22, 2021.
8. G. T. Schwartz, "The myth of the Ford Pinto case," *Rutgers Law Review*,

vol. 43, pp. 1013–1068, 1991. Available: http://www.pointoflaw.com/articles/The_Myth_of_the_Ford_Pinto_Case.pdf

9. M. Gladwell, "The engineer's lament: Two ways of thinking about automotive safety," 2015, accessed 2022-04-03. Available: https://www.newyorker.com/magazine/2015/05/04/the-engineers-lament
10. B. N. Cardozo, "Meinhard v. Salmon, 249 NY 458," 1928, Court of Appeals of New York; Accessed 2022-03-31. Available: https://www.nycourts.gov/reporter/archives/meinhard_salmon.htm
11. Deloitte, "Trustworthy AI," 2022, accessed 2022-04-04. Available: https://www2.deloitte.com/us/en/pages/deloitte-analytics/solutions/ethics-of-ai-framework.html
12. IBM, "Trustworthy AI is human-centered," 2022, accessed 2022-04-04. Available: https://www.ibm.com/watson/trustworthy-ai
13. A. Koene, C. Clifton, Y. Hatada, H. Webb, M. Patel, C. Machado, J. LaViolette, R. Richardson, and D. Reisman, "A governance framework for algorithmic accountability and transparency," European Parliamentary Research Service, Brussels, 2019.
14. K. Bryan, C. Lamoureux, and D. Lonergan, "2021 Year in review: Biometric and AI litigation," 2022, accessed 2022-07-24. Available: https://www.consumerprivacyworld.com/2022/01/2021-year-in-review-biometric-and-ai-litigation/
15. E. Liscio, M. van der Meer, L. C. Siebert, C. M. Jonker, and P. K. Murukannaiah, "Axies: Identifying and evaluating context-specific values," in *Proc. International Conference on Autonomous Agents and MultiAgent Systems*. London: IFAAMAS, 2021, pp. 1–9.
16. M. Rodriguez-Soto, M. Serramia, M. López-Sánchez, and J. A. Rodríguez-Aguilar, "Instilling moral value alignment by means of multi-objective reinforcement learning," *Ethics and Information Technology*, vol. 24, no. 1, pp. 9:1–9:17, 2022.
17. A. C. Kurtan and P. Yolum, "Assisting humans in privacy management: An agent-based approach," *Autonomous Agents and Multi-Agent Systems*, vol. 35, no. 1, pp. 7:1–7:33, 2021.
18. R. Aydoğan, Ö. Kafalı, F. Arslan, C. M. Jonker, and M. P. Singh, "NOVA: Value-based negotiation of norms," *ACM Transactions on Intelligent Systems and Technology*, vol. 12, no. 4, pp. 45:1–45:29, 2021.
19. M. Fisher, V. Mascardi, K. Y. Rozier, B.-H. Schlingloff, M. Winikoff, and N. Yorke-Smith, "Towards a framework for certification of reliable autonomous systems," *Autonomous Agents and Multi-Agent Systems*, vol. 35, no. 1, pp. 8:1–8:65, 2021.
20. J. Waldron, "Moments of carelessness and massive loss," in *Philosophical Foundations of Tort Law*, D. G. Owen, Ed. Oxford: Clarendon, 1995, ch. 17, pp. 387–408.

Amika M. Singh is a law student at Harvard Law School. Her interests include civil liberties and the use and governance of AI. She is the recipient of the Harvard Club of New York NAACP LDF Fellowship. She has completed internships at the NAACP Legal Defense and Education Fund and the American Civil Liberties Union of Georgia. Contact her at asingh@jd23.law.harvard.edu.

Munindar P. Singh is a Professor in Computer Science and a co-director of the Science of Security Lablet at NC State University. His interests include the engineering and governance of sociotechnical systems, and AI ethics. Singh is a Fellow of AAAI, AAAS, ACM, and IEEE and a former Editor-in-Chief of *IEEE Internet Computing* and *ACM Transactions on Internet Technology*. Contact him at singh@ncsu.edu.