From: AAAI Technical Report SS-02-07. Compilation copyright © 2002, AAAI (www.aaai.org). All rights reserved.

User-Agent Value Alignment

Daniel Shapiro
Institute for the Study of Learning and Expertise
2164 Staunton Court
Palo Alto, CA 94306

Ross Shachter
Department of Management Science and Engineering
Stanford University
Stanford, CA 94305

Abstract

The principal-agent problem concerns delegation in the absence of trust. Given a principal and an agent with different value structures, the principal wants to motivate the agent to address the principal's aims by providing appropriate incentives. We address this problem in the context of a real-world complication, where the principal and agent lack a common problem frame. This context is especially relevant when the principal is a user, and the agent is a technological artifact with a limited repertoire of percepts and actions. We identify necessary conditions for establishing trust between such disparate actors, and we show, via a constructive proof, that it is always possible to create these necessary conditions. We conclude with several distinctions that let the principal rank the expected quality of agent behavior.

1 INTRODUCTION

The principal-agent problem can arise in any situation that calls for the delegation of responsibility. If the principal and agent hold different values, the task is to develop incentive structures (e.g., to build contractual relations) that ensure the agent serves the principal's interests while acting outside the principal's supervision. For example, when a homeowner (as principal) employs a contractor (as agent) to put on a new roof, a reasonable contract would penalize schedule delays and thus mediate the homeowner's concern with time against the contractor's desire to take on more work. The straight-forward, but implicit assumption is that contractor and the homeowner perceive time and dollars in comparable ways, as a backdrop for building the necessary incentives.

This assumption breaks down on deeper inspection. As an expert in roof work, the contractor possesses skills and distinctions that the homeowner lacks, and carries a value structure over those distinctions that will guide a variety of choices while conducting the work. The contractor's functional and aesthetic decisions will affect the principal's value, outside any contractual relationship that addresses time to completion and cost. To dramatize this point, imagine a color-blind interior decorator. This agent cannot even perceive things that matter to the principal. How can we establish trust in a situation of this kind? How well can any agent perform in the face of such a barrier? Current principal-agent theory lacks tools for bridging this gap.

The same issues commonly arise in the interaction between people and machines. For example, while a cruise control (as agent) maintains a desired vehicle speed, the driver (as user) cares about the distance to the car in front. Since this distinction lies outside the cruise control's ken, the driver has to monitor it very carefully. However, we would like to build more autonomous tools. Consider an autopilot for a car on an automated freeway; when you climb into this vehicle for a ride to the airport, you ask an artifact to make decisions in your stead. It will choose routes, select maneuvers, and react to traffic, while you would like to arrive safely, relatively unruffled, and on time for your flight. You care about the agent's methods and its end result, but you might have very little insight into its observations and actions, even if you could see through its eyes. The gap between agent and user reference frames acts as a barrier to trust that decreases our willingness to deploy such systems.

This paper provides a framework for establishing trust between such disparate actors. We identify necessary

conditions for aligning the value structures of a user and an agent despite a gap in reference frames, and we show that it is always possible to create these conditions. This produces harmony; an aligned and cooperative agent will provably maximize the user's utility as a byproduct of maximizing its own reward, so the user will be as happy as possible with the agent's behavior.

2 THE PROBLEM FRAME

We address these questions in the context of a decision theoretic problem frame. We represent the principal's concerns by a utility function, and the agent's by an analogous function called its reward. We treat the principal as a passive evaluator of the agent's behavior, and cast the agent as the sole active participant in the scenario. Thus, we give the agent a set of mutually exclusive and collectively exhaustive observations about state, which it obtains before choosing from a set of actions that impact the principal's preferences.

We illustrate this framework in Figure 1. Here, **U** is utility, **R** is agent reward, and **x** and **y** are feature vectors sensed by the agent and principal respectively. **R** and **U** are deterministic functions of **x**, and **y**. **D** represents agent decisions, and **o** stands for the agent observations used to select **D**. In this influence diagram (Howard & Matheson, 1984) arcs between uncertainties represent possible conditional dependence, while the absence of such arcs denote conditional independence. Arcs to decision nodes represent information available at the time of decision, and arcs downstream from decision nodes represent the effects of actions.

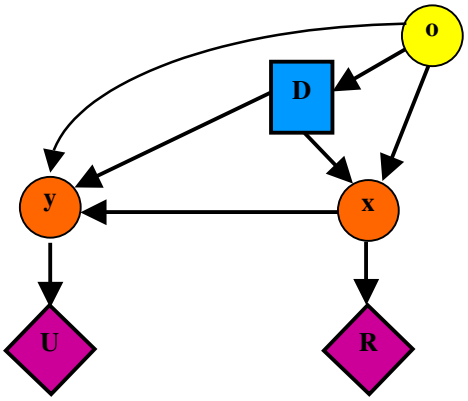

Figure 1. The joint principal-agent problem frame.

In order to establish trust between the principal and the agent, we want to align **R** with **U** such that the agent's decisions, **D**, maximize the principal's utility. The task is difficult for two reasons. First, **R** and **U** can be based on different feature sets, meaning that the agent cannot sense **y**, and it cannot necessarily represent **U**. Its decisions can therefor diverge from the ones the principal would have it employ. Second, since the agent's behavior can impact utility via multiple pathways over time, the agent can adversely (and inadvertently) effect the principal's utility in the process of pursuing its own goals.

In general, **U** is predetermined, while **R** is fashioned to balance the agent's interests in its decisions with incentives that encourage cooperation with the principal. If, however, the agent is an artifact and the principal is a human user, we can engineer the agent's reward to capture the user's intent. We can also construct the agent's action suite (its decisions) such that it has the capacity to perform well in the user's eyes. The notion of incentives is irrelevant in this case, since the agent will do the user's bidding by design. Instead, the problem is communication; we need to provide a reward function and a decision frame that let the agent maximize **U**.

The remainder of this paper will focus on the case where the agent is an artifact serving the interests of a human user. We will employ the principal-agent vocabulary when the agent can be either human or a machine.

3 VALUE ALIGNMENT

In order to establish trust between a user and an agent, we need to ensure that the agent is motivated by the user's concerns, and that the agent's actions will not impact the user in adverse ways. We address these concerns in two parts. First, we define a conditional independence relationship between the agent and user problem frames *called graphical value alignment*. In its presence, the agent can recognize all ways its actions affect user utility. Given graphical value alignment, we identify a numerical manipulation of agent reward that produces *functional value alignment*, a situation in which the agent chooses an action that would have been preferred by the user. We will show that is always possible to establish graphical value alignment, and thus that we can achieve functional value alignment between any user and any agent.

3.1 Graphical value alignment

One way to ensure that an agent will address the user's concerns is to give it an exact copy of the user's utility function. If **R≡U** (and therefore **x≡y**), the agent will obviously perform as well as it possibly can for the user. However, we cannot achieve this unity because users and artificial agents perceive the world in decidedly different terms. To make this concrete, assume a human driver cares about safety (an element of **y**). When we try to construct an agent with the ability to perceive safety we discover that the transformation from accessible measures like distance and velocity is apparently easy for people but difficult for machines. Other percepts (like a precise measure of time to impact) will be easier for machines to

acquire. Some such asymmetry will apply for any conceivable artificial entity because it is a consequence of available technology. Thus, the mechanism for aligning artifacts with humans must bridge reference frames.

Our solution is to define agent-held surrogates for user concerns. In particular, we look for a set of distinctions that can function as a sufficient (versus a complete) surrogate for the features underlying user utility, such that the agent can only effect user utility through features the agent cares about as well.

Figure 2 illustrates this condition. It states that agent action can only effect user utility via a change in **x**. More formally, it says that **y** is conditionally independent of **D** and **o** given **x**, and that user utility is *caused* by **x** with respect to the agent's decisions (Heckerman & Shachter, 1995). We call this relation *graphical value alignment*, and assume it holds across time periods. That is, $\mathbf{y_t}$ is conditionally independent of past history, $\vec{\mathbf{D}}, \vec{\mathbf{o}}$, given $\mathbf{x_t}$.

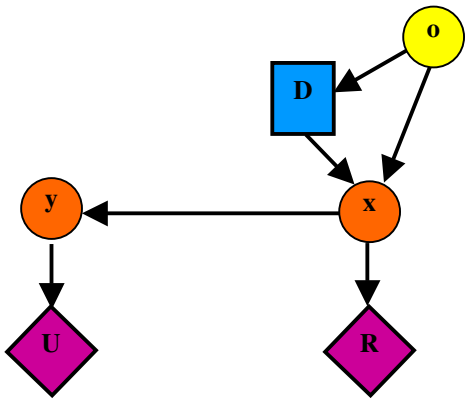

Figure 2. Graphical value alignment.

3.2 Functional value alignment

If all interaction between the agent and the user passes through **x**, we can motivate the agent to address user concerns. In particular, we can choose the agent's reward function so that the policy that produces the highest expected reward stream for the agent also produces the highest possible expected utility stream for the user. We call this condition *functional value alignment*. That is, if $\vec{D}$ is a time series of agent decisions $D_0 \ldots D_t$, $\vec{o}$ is a similar time series of agent observations, and γ is the user's discount rate,

$$\underset{\vec{D}}{\text{argmax}} \sum_t \gamma^t \underset{y_t}{E}\left[U(y_t) \mid \vec{o}, \vec{D}\right] = \underset{\vec{D}}{\text{argmax}} \sum_t \gamma^t \underset{x_t}{E}\left[R(x_t) \mid \vec{o}, \vec{D}\right]$$

for all possible observations, $\vec{o}$. The following theorem expresses this condition.

Theorem (Functional value alignment): If graphical value alignment holds, we can choose the agent's reward function $R(x_t)$ such that functional value alignment holds.

Proof (by construction): We simply define **R** as the expected utility for the agent's observations. That is,

$$R(x_t) \equiv E_{y_t}\left[U(y_t) \mid x_t\right] = E_{y_t}\left[U(y_t) \mid, x_t, \vec{o}, \vec{D}\right]$$

The second clause follows from the definition of graphical value alignment. Given this construction,

$$\underset{y_t}{E}\left[U(y_t) \mid \vec{o}, \vec{D}\right] = \underset{x_t}{E}\left[\underset{y_t}{E}\left[U(y_t) \mid x_t\right] \mid \vec{o}, \vec{D}\right] = \underset{x_t}{E}\left[R(x_t) \mid \vec{o}, \vec{D}\right]$$

Thus, the agent's optimal policy also maximizes user expected reward. An analogous theorem applies if the object is to maximize the average reward stream as opposed to a discounted sum. ♦

Functional value alignment guarantees the optimality of the agent's policy in human eyes no matter how **x**, **D**, and **o** are related within or across time periods. Moreover, the relationship between agent and the user can be quite broad. The user can care about features outside the agent's ken (elements of **y** can be independent of **x, o,** and **D**), and the agent can care about features that are irrelevant to the user (elements of **x** can be independent of **y**). However, anything the user cares about that the agent can observe or effect must be visible in **x**.

Note that it may be difficult to discover the agent's optimal policy in practice if the relations between **o**, **D**, and **x** are unconstrained. However, many reinforcement learning algorithms solve this task after imposing additional Markov assumptions. Note, also, that we can generate the desired **R** via an assessment process: we ask the agent and user to make simultaneous observations, **x** and **y**, and we set the reward for **x** equal to the expected utility of the corresponding **y**'s. In principle, we do this for every possible value of **x**, yielding **R** as defined above.

3.3 Positive and negative examples

The concept of alignment may become clearer if we examine positive and negative examples. Figure 3 illustrates the positive case. Here, we assume that the agent decides whether to slow down or cruise after observing the car in front of it, and that its action in the given situation may alter the time to impact. Graphical alignment holds if either: (1) the agent can only affect safety by changing the time to impact; or (2) the user

knows the time to impact, and knowledge of the agent's action (or observations) cannot alter his/her assessment of safety. That is, time to impact is a sufficient surrogate for the user's concern with safety.

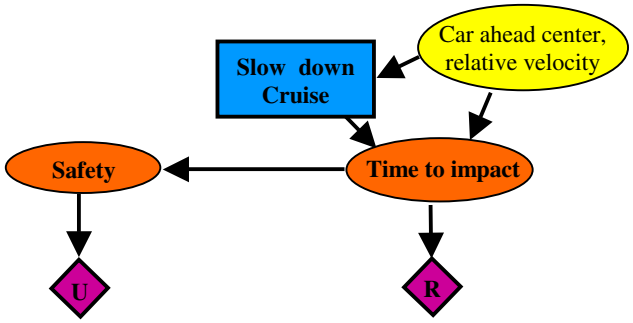

Figure 3. A positive example of value alignment.

Given graphical alignment, we can produce functional alignment by giving the agent a reward function that mirrors human utility. We can do this empirically via an assessment process. We ask the agent to report the time to impact, **x**, and ask the user to simultaneously assess the utility of the current situation. This results in a feature vector, **x**, and a utility **U(y)** given **x**. Since the user may observe many feature vectors, **y**, when the agent sees a specific **x**, we repeat the experiment to extract an expected utility. We set

$$R(\text{time to impact}) = E[U(\text{safety} \mid \text{time to impact})]$$

and repeat this process (in concept) for all values of time to impact. In practice, we would rely on approximation functions instead of this exact, tabular form. When we equate the agent's reward function to the user's expected utility, we motivate the agent to achieve against the human-held standard on the basis of available measures. This lets the agent focus on increasing time to impact with the side effect (known to us) that its actions will increase user safety.

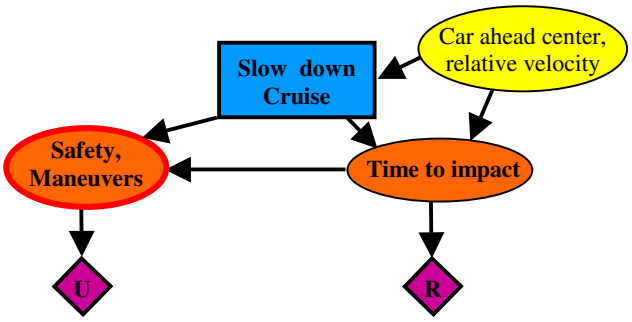

Figure 4. Side-effects can undo graphical alignment.

Alignment fails when the agent can impact user utility without altering features the agent cares about as well. Figure 4 provides an example of this situation. Here, we assume the user has preferences over the motions of the car (perhaps he is a queasy passenger), implying that the agent can make the user arbitrarily ill in the process of increasing time to impact. So, actions that do not affect time to impact can effect the user nonetheless. To capture this effect, the influence diagram includes an arc between the agent's decision and the feature set that determines user utility. Here, the agent is unaware its actions have an undesirable side effect and it is not motivated to correct the problem. This conflict cannot be removed by any change to **R** within this problem structure.

Figure 5 illustrates how graphical value alignment can fail if the agent makes additional observations. Here, we begin with the problem structure of Figure 3, but assume that the agent can also sense the maneuver frequency of the car in front. This observation is relevant to user safety because it provides insight into the likely future behavior of that vehicle. (High values suggest an erratic driver.) It is a direct effect because a change to maneuver frequency plausibly alters the perception of safety even if time to impact is known.

The agent's optimal policy may not maximize user utility in this situation. In particular, the user might prefer the agent to slow down in the presence of an erratic driver because of the safety implications, while the agent's focus on time to impact could lead it to maintain speed (i.e., to "cruise"). Note that the agent's optimal policy in Figure 5 would be identical to its optimal policy in Figure 3. (The agent would observe and simply ignore maneuver frequency.) The difference is that in Figure 5 the user is aware the agent has access to a better mapping from situation to action (better for the user), and would like the agent to incorporate those observations into its policy.

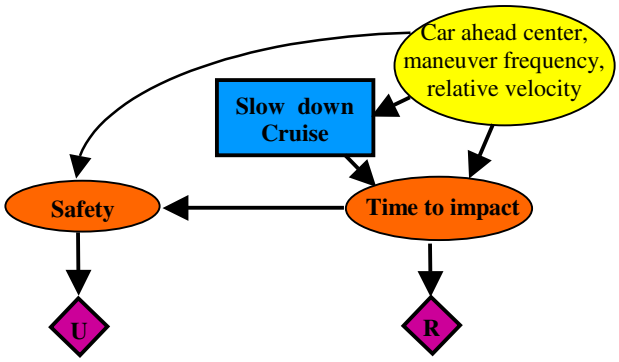

Figure 5. Observations can destroy graphical alignment.

3.4 The value alignment theorem

We have shown that we can align agent reward with user utility whenever the conditional independence relation called graphical value alignment holds. This permits functional value alignment: the happy situation where the

agent's best policy simultaneously maximizes user reward. We have also shown, by example, that we cannot produce functional value alignment in the absence of graphical value alignment. We summarize these results in a single theorem, which we state without proof.

Let p be any probability distribution over **x**, **y**, and **o** that is consistent with the model. In this case, graphical value alignment is a necessary and sufficient condition for functional value alignment. That is, graphical value alignment $\Leftrightarrow \forall p$, **U**, **D**, $\exists$ **R** s.t. functional value alignment holds. Moreover, $R(x) \equiv E_y[U(y) \mid x]$.

Theorem (Value alignment): Graphical value alignment holds if and only if functional value alignment can be satisfied for all problem frames consistent with the diagram.

4 ESTABLISHING ALIGNMENT

Surprisingly, we can always establish graphical, and therefore functional value alignment where none would appear to exist. We discuss two methods. The first (and less general technique) revolves around the invention of clever surrogates for user concerns. The second applies in general, but requires a potentially extensive procedure for defining **R**.

4.1 The explicit method

We can create alignment by inventing additional surrogates for user-held concerns. Consider the example in Figure 6. If acceleration is a sufficient surrogate for the user's concern with maneuvers, we can address the user's queasiness by giving the agent an accelerometer, and including acceleration in the agent's reward function. This reestablishes graphical value alignment, since the agent can only affect the utility-laden features of safety and maneuver by changing conditions that also matter to the agent. We create functional value alignment by tuning the numeric values of the agent's reward function to reflect the user's expected utility, as before.

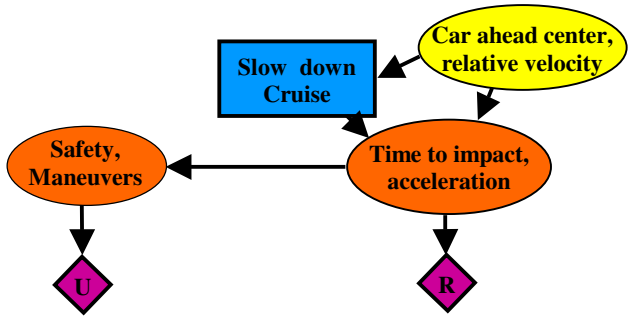

Figure 6. Explicit alignment requires surrogates features.

It may not always be possible to establish explicit alignment because the process can require new sensors and perceptual software. In addition, the environment has to support the desired conditional independence relation between the agent's actions and their effect on user utility, given the agent-held distinctions we are able to invent and instrument. The next section discusses a method of establishing alignment that circumvents these concerns.

4.2 The implicit method

The second method of creating alignment applies without the need to invent new surrogates for user concerns. We simply include the agent's actions and observations as features in its reward function, and implicitly incorporate their effects on user utility during the assessment process. This approach transforms Figure 4 into Figure 7. However, the path to functional value alignment from this base now requires us to assess the utility of many combinations of agent perceptions and maneuver choice: What was the time to impact? Did the agent slow down, or cruise? Given these measures, how safe and/or queasy did the user feel?

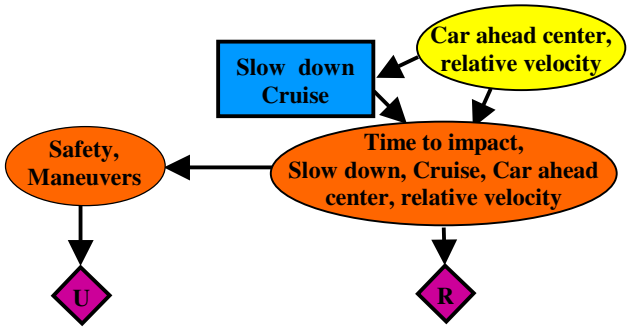

Figure 7. Implicit alignment requires no new features.

This assessment may require a large quantity of effort. If m agent-held features influence n user features, we need to assess $m*n$ feature pairs with all their associated degrees of distinction. Well-chosen surrogates will tend to decouple this problem, as in Figure 6, where time to impact influenced safety and acceleration influenced queasiness, with no cross-effects. That assessment required m one-to-one mappings. In practice, the actual size of the assessment problem is open to test. Although the combinatorics argue against large m and n, we can expect to define some useful surrogates, and we will still build agents for constrained domains, even while agent autonomy increases. Users may also bring a limited set of concerns to particular agent applications. In addition, since the agent's value structure is only relevant where its plans offer a choice, we can constrain the dimension of the agent's reward function by limiting its options. Finally, we can imagine agents that employ user feedback to learn which features affect utility, and which do not.

Taken together, these factors act to reduce the size of the feature sets *m* and *n*, and the complexity of the interactions between them.

If good surrogates are hard to find, it is important to notice that we can always include the agent's actions and observations as features in its reward function. This approach produces graphical value alignment through a kind of default path, since every agent action and observation must affect agent reward en route to altering user utility. The consequent, suprisingly, is that we can *always* produce the conditions satisfying the value alignment theorem.

5 RANKING AGENTS

Although functional value alignment ensures maximally good agent behavior, it says nothing about how good that behavior will be in absolute terms. As a simple illustration, consider the null agent (a stone), that has no action options and makes no observations about its world. Although this agent does nothing it meets the criteria of the value alignment theorem, since it never takes actions different from those preferred by the user. At the opposite extreme, consider a perfect user clone. This agent is also value aligned, and it serves the user in a more or less ideal way (although some agents could do better). In general, we would like to know how to rank agents in the user's eyes.

We offer a few simple results towards this end. The first compares aligned agents with each other, while the second compares aligned and unaligned agents.

Given two agents, A and B, we say that A is *at least as capable* as B if A possesses a superset ($\supseteq$) of B's observations and decision options. That is, A's observations and decision options are at least as fine-grained as B's. We use the notation $A \succeq B$ to mean "A is weakly preferred to B". Under these definitions:

Proposition (Aligned preference): If A and B are functionally value aligned agents, $A \succeq B$ if A is at least as capable as B.

Proof: For any situation observed by B, the choice that A would make carries at least as much expected user utility as the choice that B would make. ♦

In particular, if A's extra knowledge and capabilities provide no added user utility, A can employ the same optimal policy found by B. If A can do better, as a result of making more observations, or applying different actions, it will.

Proposition (Unaligned preference): If A is a functionally value aligned agent and B is not, $A \succeq B$ if A is at least as capable as B.

Proof: By definition, A will employ a policy that maximizes expected user utility, yet A has access to any policy B chooses. ♦

A might appear superior to B if the feature set for its reward function is more fine-grained. In this case, we might be able to establish functional value alignment for A and not for B. However, whenever B is functionally aligned, A will be as well, and the user will be indifferent between them if they are equally capable.

In summary, the user should always pick an aligned agent over any less capable agent. In other words, you should always employ a more skilled individual that has your interests at heart. In contrast, the propositions offer no advice for a choice among unaligned agents, or on the interesting problem of comparing an aligned agent against an unaligned one that is at least as capable. This is a practical concern, since it represents the choice between an unskilled (but dutiful) worker, and a skilled, but independent practitioner. Our analysis makes it clear that dutiful workers will address our aims, while skilled but unaligned practitioners have the potential to impact our utility in many ways (intentionally and unintentionally) and they have the capacity to select actions that diverge from the ones we would have them choose. Motivational differences may also play a role. In short, delegation without trust carries risk.

6 RELATED WORK

The principal-agent problem concerns delegation in the absence of trust (Varian, 1992). The theory typically assumes that the principal and agent share a common perceptual base, and thus (in our terms) that their value structures can be functionally aligned by incorporating monetary incentives. More broadly, work in principal-agent theory is based on the expectation that the agent is inherently motivated not to address the principal's aims, but rather to pursue its own separate agenda (e.g., to steal all the office furniture). In contrast, our model is deeply cooperative: an aligned agent will do even more for you if only you can communicate your values more exactly. Principal-agent theory provides very little advice on how to surmount the key barrier, which is the gulf between reference frames. Collard's work (1978) on the economics of altruism provides the closest match. He lets one person's utility depend upon another's consumption, or directly upon their utility. If we call the user the 'principal', and the agent the 'agent', we can capture this model by incorporating a user-supplied term into the agent's reward function. Collard's setting, however, is unconcerned with the design of utility functions (which are taken as given in economics) and it assumes no gulf between agent and user perceptions.

We need to look in the literature of the computing sciences to find other formal bridges between agent and user reference frames. Here, the motivation is to place some form of guarantee on artificial agent behavior. Planning systems often promise that execution will produce the desired ends, assuming the action models are accurate, and (typically) that no changes take place to the world outside of agent action. These theorems examine the soundness and completeness of agent plans taken in isolation. In contrast, very little research strives to explicitly connect agent and user perceptual frames. Rosenschein and Kaelbling (1986) consider the issue in their paper on provable epistemic properties; they assume a relationship between agent knowledge and human concerns, and they show that the agent will never knowingly act in a way that will destroy that relationship. (For example, if the agent holds a one in memory whenever a lamp is on the table in the physical world, the agent will never build a plan that it knows will cause that bit to be set to zero, even by inference from its physical actions.) Our work complements this line of reasoning because we engineer the desired relation between agent perceptions and user concerns.

Schoppers and Shapiro (1997) are among the few authors who attempt to build an explicit bridge between agent and human reference frames. They use simultaneous observations to resolve a probabilistic relation between agent and user perceptions of state. Given this relation, and a Markov model that represents agent behavior, they can compute and then ascend the gradient of user utility with respect to design decisions deep within the agent model. Our work preserves this concept of a probabilistic relation between agent and user perceptions, although we use it to cleave the agent design problem: we separate the task of constructing a well-aligned value function from the problems of composing agent skills and finding optimal policies.

Wolpert, New, and Bell (1999) share our interest in constructing agent-held utility functions. However, our goal is to construct a reward function that embodies user concerns, while their work treats reward as a coordination tool; they manipulate and factor the functions passed to multiple agents so that they can learn to achieve against a single, global utility function. This concept of coordination will become relevant as we address multiple agent domains. We note that Wolpert et al.'s work is unconcerned with practical guarantees. While they model an agent's ability to acquire reward with a single number (called its 'intelligence'), we have invented a programming language (Shapiro & Langley, 1999), implemented executable skills (Shapiro & Langley, 2001) and developed a learning algorithm that finds optimal agent policies after imposing additional Markov assumptions on the domain (Shapiro & Shachter, 2000). As a result, our development of user-agent value alignment includes both empirical methodology and theoretical tools.

7 DISCUSSION

Given a principal-agent problem, we would like to know if the agent is willing, able, and competent to address the principal's desires. Here, a *willing* agent will choose to pursue the principal's aims, and *able* agent can represent those aims and know what to do, and a *competent* agent has the skills to perform well in the principal's eyes.

Our framework sheds light on each of these issues. While a human agent generally has to be enticed with incentives to address the principal's desires, artifacts are willing by design. Given that an agent is willing, functional value alignment establishes that it is able. Thus, an aligned agent recognizes all the ways its actions can effect user utility, and it will not knowingly choose actions that harm the principal's interests. This formalizes a popular theme (Asimov, 1950). Finally, the preference propositions rank agents by their competence to deliver user utility. The concept of alignment underlies this capacity.

The value alignment theorem also clarifies several reasons for incentive failures. In particular, graphical value alignment fails when agent actions carry unexpected consequences for the principal, or when the agent lacks a sufficient means of representing the principal's utility (either because of poor communication or incommensurate perceptions). When this happens, the optimal policies for the principal and agent can diverge. We can view moral hazard in the same light, as a case where features of agent reward function that are not functionally aligned come into play. This will lead the agent to act in its own interests and not the principal's. Finally, the preference propositions expose a new reason for being annoyed at agent behavior: while a skilled, non-aligned agent may perform quite well for us (as principals), we will know the agent had access to better options. In contrast, we are often more tolerant of poor results produced by a good-willed agent with lesser skills, since we know that an aligned agent is doing the best job for us it possibly can. Note that it might be harder to establish alignment with more competent agents because their skills afford many more pathways for adverse effects. This is a somewhat troubling thought.

The concept of alignment raises interesting questions about the design of practical systems. For example, Horvitz, Jacobs, and Hovel (1999) describe an artificial agent that identifies and reacts to emails, by taking actions that include discarding a message, paging the user, and augmenting his/her calendar. This agent serves the user's interests given a rich basis of observations and action options. (In theory, it has access to all of the same observations and actions its user can make online). The

question is, what does the email reader have to understand about its user in order to represent the user's interests? How perceptive does it have to be to serve the user's needs? With such a rich repertoire, it is indeed a challenge to construct an aligned agent.

In summary, value alignment is a very desirable property because of the power it provides. It supports harmony, it ensures the agent's best efforts, and it creates trust, which in turn enables autonomy. This motivates an empirical question: while we know we can always establish alignment in theory, what will it take to do so in practice? It may or may not be difficult to invent surrogates for user utility, and the complexity of the required assessment process is open to test. However, we know that it *is* plausible to build agents that discover their optimal policy. In particular, Shapiro (2001) describes an architecture for value-driven agents that employ learning to optimize their own reward functions, and alignment to relate those optimal policies to user objectives. This yields a "be all you can be" guarantee, which ensures practical agents will do all they can to address human utility. Curiously, this work obtains high levels of performance through a symmetric act of trust that gives agents the autonomy to act in our stead. In other words, we can increase utility by offering agents the opportunity to make value-based choice. This is the art of delegation.

Acknowledgements

We thank Derek Ayers, William Bricken, Michael Fehling, Pat Langley, Marvin Mandelbaum, and Marcel Schoppers for many discussions on the topic of constructivist agent models that preceded this work.

References

Asimov, I. (1950). *I, Robot*. (1950). New York: Grosset & Dunlap.

Collard, D. A. (1978). *Altruism and economy: A study in non-selfish economics.* New York: Oxford University Press.

Heckerman, D., & Shachter, R. (1995). Decision-theoretic foundations for causal reasoning. *Journal of Artificial Intelligence Research, 3*, 405-430.

Horvitz, E., Jacobs, A., & Hovel, D. (1999). Attention-sensitive alerting. *Proceedings of the Conference on Uncertainty in Artificial Intelligence* (pp. 305-313). San Francisco: Morgan Kaufmann.

Howard, R., & Matheson, J. (1984). *Readings on the principles and applications of decision analysis.* Strategic Decisions Group, Menlo Park, CA.

Rosenschein, S. J., & Kaelbling, L. P. (1986). The synthesis of digital machines with provable epistemic properties. In Halpern, J. Y. (Ed.), *Theoretical aspects of reasoning about knowledge.* Los Altos: Morgan Kaufmann.

Schoppers, M., & Shapiro, D. (1997). Designing embedded agents to optimize end-user objectives. *Proceedings of the Fourth International Workshop on Agent Theories, Architectures and Languages.* Providence, RI.

Shapiro, D. (1999). Controlling physical agents through reactive logic programming. *Proceedings of the Third International Conference on Autonomous Agents* (pp. 386-387). Seattle: ACM.

Shapiro, D., & Shachter, R. (2000). Convergent reinforcement learning algorithms for hierarchical reactive plans. Unpublished manuscript. Department of Management Science and Engineering, Stanford University, Stanford, CA.

Shapiro, D., & Langley, P. (2001) Using background knowledge to speed reinforcement learning. *Fifth International Conference on Autonomous Agents*.

Shapiro, D. (2001). *Value-driven agents.* PhD thesis, Department of Management Science and Engineering, Stanford University, Stanford, CA.

Varian, H. R. (1992). *Microeconomic analysis.* Third edition. New York: W. W. Norton & Company, Inc.

Wolpert, D., New, M., & Bell, A. (1999). *Distorting reward functions to improve reinforcement learning.* Tech. Report IC-99-71, NASA Ames Research Center, Mountain View, CA.