Designing Preferences, Beliefs, and Identities for Artificial Intelligence

Vincent Conitzer
Department of Computer Science
Duke University
Durham, NC 27708, USA

Abstract

Research in artificial intelligence, as well as in economics and other related fields, generally proceeds from the premise that each agent has a well-defined *identity*, well-defined *preferences* over outcomes, and well-defined *beliefs* about the world. However, as we design AI systems, we in fact need to *specify* where the boundaries between one agent and another in the system lie, what objective functions these agents aim to maximize, and to some extent even what belief formation processes they use.

The premise of this paper is that as AI is being broadly deployed in the world, we need well-founded theories of, and methodologies and algorithms for, how to design preferences, identities, and beliefs. This paper lays out an approach to address these problems from a rigorous foundation in decision theory, game theory, social choice theory, and the algorithmic and computational aspects of these fields.

Problem Overview

Agents are generally assumed to have a well-defined *identity* over time, well-defined *beliefs* about the world as it is and how it will develop over time, and well-defined *preferences* over the different ways in which things may proceed. Perhaps the main exception is that in machine learning, in fact, we do develop techniques for *obtaining* beliefs about the world and how it develops, but this is not always integrated into the more decision- and game-theoretic work on AI.[1] But the agent's preferences (or the *objective* it pursues) are usually taken to be given exogenously. Economic theory, and the AI literature that is based on the same ideas, proceeds from the maxim *de gustibus non est disputandum*—there is (to be) no arguing about taste. Similarly, the agent's identity is usually taken to be clear; occasionally, in economics, there is some discussion of whether, for example, we can reasonably consider a household to be a single agent or we need to split it up into its individual members, but in the end most of the analysis focuses on other aspects. Fundamental questions of how we should think about identity are generally left to philosophers (e.g., the ship of Theseus: does an object that has had all its parts replaced remain the same object?), and the same is true for questions of what preferences we *should* have (identifying "the good life").

The premise of the research vision layed out here is that as AI is getting broadly deployed in the world, we, as AI researchers, do in fact need to address these types of questions. In earlier stages of AI research, this was naturally not the top priority: it was hard enough to get AI systems to behave effectively *even if* their identities, beliefs, and preferences were clearly specified. Techniques were (and of course continue to be) tested in abstract domains. For well-specified games such as chess, Go, or poker, the identities, beliefs, and preferences of the agents are in principle clear from the definition of the game. In contrast, when systems are deployed in messy, ambiguous, real-world environments, the pursuit of what may at first appear to be natural objective functions can easily lead us astray. In deployed machine learning systems, maximizing the fraction of instances classified correctly may result in systems that discriminate against certain subgroups of the population. In social media, showing posts to maximize the amount of positive feedback may result in a more polarized society. In both cases, the solution to the problem may not be as simple as a straightforward modification of the objective function; it may also require fundamentally rethinking the modeling of the situation, including on whose behalf different parts of the AI system are supposed to act.

As can be seen even from these two examples (which certainly do not constitute an exhaustive list), the precise issues that need to be addressed differ from domain to domain. This suggests that there is much valuable domain-specific research to be done on these topics. However, it is also worth taking a step back and attempting to identify general principles. Generally speaking, are we thinking in the right way about how to specify preferences (or objectives) for a system, about defining the boundaries of agents (e.g., should the system be split up into multiple agents that represent different people or other real-world stakeholders or objectives), and even about how exactly these agents form beliefs? These themes are appearing in various conversations among researchers with an interest in AI. At one extreme, there is a community of people that worry about how to specify goals for hypothetical superintelligent AI (artificial *general* intelligence that broadly exceeds human capabili-

[1] There are, of course, clear exceptions, including some of the work on learning in games; for an overview, see Fudenberg and Levine (1998) or the 2007 issue of the journal Artificial Intelligence on multiagent learning.

ties) in such a way that the outcome will be good for humanity; this community also ponders whether such superintelligent AI would eventually be unipolar or multipolar (see, e.g., Bostrom (2014); Tegmark (2017)). These are intriguing thought experiments, but most of the proposed approaches are laid out at a very high level and it is not immediately clear how best to evaluate them, which is perhaps inevitable given the speculative nature of the subject matter. On the other hand, there are also more concrete recent proposals for how to specify AI systems' preferences (and/or decision methodologies) that can be evaluated in the absence of superintelligent AI systems. For example, Abel, MacGlashan, and Littman (2016) argue for a reinforcement learning approach to these topics, and Hadfield-Menell et al. (2017) argue that we should design AI agents to pursue an objective function that is not directly given to them.

This paper lays out a research agenda to rigorously investigate and evaluate how to specify preferences (and decision methodologies), identities, and beliefs of artificially intelligent agents. The approach is distinguished from prior approaches by being based primarily on decision theory, game theory, and social choice theory.

Aggregating Multiple Signals into Consistent Preferences

When the stakes are high in the design of an agent's preferences, we will likely rely on more than one assessment of what those preferences ought to be. For example, we may want our agent's preferences to reflect the preferences of multiple stakeholders. If so, we may wish to sample some stakeholders, elicit what each of them thinks should be the objective function according to which the agent operates, and then aggregate these multiple objective functions into a single consistent objective function for the agent. This turns the problem into a *social choice* problem; the link to social choice in such a context has already been observed several times (Greene et al. 2016; Conitzer et al. 2017; Noothigattu et al. 2018; Zhang and Conitzer 2019). This general approach has already been used in the context of algorithms for finding matchings in *kidney exchanges* (Freedman et al. 2018) as well as in the context of self-driving cars making emergency decisions (Noothigattu et al. 2018).

What is the best way to aggregate multiple objective functions into a single one? It appears that the traditional models used in social choice theory are not ideal for this problem, though they certainly provide insight. If each individual provided a *ranking* of all the available alternatives, this would fit perfectly in the standard model of voting theory. But this does not scale: we want the resulting AI system to operate autonomously in a world in which it can encounter any of exponentially many possible scenarios (self-driving car), search through spaces with exponentially many possible alternatives (kidney exchange), etc. We cannot elicit a ranking of exponentially many alternatives from a human being, so we have to approach the problem somewhat differently.[2]

[2]We *can* learn a model that *predicts* how someone would rank any pair of alternatives; see, e.g., Noothigattu et al. (2018). Still, such a model must be more concise than an arbitrary ranking.

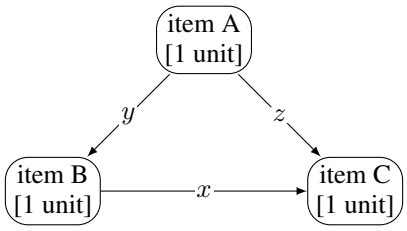

Figure 1: Weighted graph representing the numerical tradeoffs chosen by an individual voter. An arrow from one item to another with weight w represents that one unit of the former is considered as valuable as w units of the latter. The tradeoffs are *consistent* if $z = x \cdot y$.

A natural approach is to assume *linearity* in the objective function: e.g., assuming that the value gained from saving patients A and B is the sum of the values of saving each individually. Under this assumption, a first approach may be to ask each voter for the value of saving a patient of type A, etc. But it is not clear in which units to express this. Instead, it is more natural to ask for *relative* judgments: e.g., a given individual may consider saving 100 patients of type A to be as valuable as saving 103 of type B. How should we aggregate such relative judgments?

As it turns out, a framework that we developed earlier in a different context (Conitzer, Brill, and Freeman 2015; Conitzer et al. 2016) fits this setup quite well. In this framework (changing the language to better fit the present context), we receive multiple *votes*, where each vote provides an assessment of the relative values of each pair of *items*. (In the above, awkwardly, an "item" would be a patient of a particular type, but at least the word "item" makes it clear that these could be all sorts of entities.) Figure 1 exemplifies such a vote. While this framework has significant limitations, it provides a natural starting point for developing more expressive frameworks that can accommodate richer ethical theories—for example, by attempting to drop the linearity assumption.

Designing Agents' Identities

Often, AI techniques are merely sprinkled into a larger system (the proverbial "raisins in the raisin bread"). Even when AI plays a dominant role in the system, the boundaries of the AI system may not be clear. For example, a self-driving car may periodically have its software updated together with all other vehicles of the same make, perhaps based on additional data recorded from some of these other vehicles. If so, it is awkward to think of each vehicle as containing its own separate AI agent; in many ways, they all just form part of a larger system that they benefit from and contribute towards.

One may conclude that there is no reason to define clearly delimited agents within the broader system. Still, the intuition that "my car has its own AI" may be onto something, if for no other reason than that it may reflect something about how the end users of the technology would *expect* it to behave. A car's owner may expect it to prioritize her safety over that of others, or to safeguard her privacy by not shar-

ing her location data to the larger system, perhaps due to a worry about whether it keeps the data sufficiently secure. Failing to satisfy these user expectations may result in decreased adoption of the technology, thereby perhaps reducing welfare overall. In the context of self-driving cars, indeed, Bonnefon, Shariff, and Rahwan (2016) have argued that people would be unlikely to buy vehicles that did not put their occupants' safety first, so that offering "selfish" vehicles may paradoxically reduce overall traffic fatalities more, by increasing adoption of a safer technology.

If we buy into this view, it leads to the question of *identity design*: what is the best way to determine where one agent begins and another ends? Should there be a separate agent corresponding to each user, representing her interests? Even this does not completely settle matters. For one, parts of the broader system may yet be housed "centrally," not representing any user in particular; how much of the system should be centralized in this sense? Also, the boundaries need not be the same for each aspect. We can have cars make selfish decisions (prioritizing their occupants) but share all data, or have them make altruistic decisions but limit data sharing.

Imperfect Recall

One key issue in the context of agent identity design is what we let the agents *recall*. In game theory, we often assume *perfect recall*, i.e., every agent always remembers everything that she knew earlier in the game. In contrast, games of *imperfect recall* specify exactly what an agent remembers and forgets at each point of the game. In games that are played by human subjects, it is generally impossible (or morally wrong) to make the human players forget specific information, which is part of the reason that these games have generally received less attention from economists.[3] In contrast, for AI systems, the possibilities for specifying what is recalled are effectively unlimited, ranging from systems that immediately forget everything to ones that instantly share among all their nodes all data about all users everywhere. This brings the concept of imperfect recall to the fore. Indeed, imperfect recall is already used in the design of poker-playing AI, to improve scalability (Waugh et al. 2009; Lanctot et al. 2012; Kroer and Sandholm 2016). But there are reasons to use imperfect recall other than scalability. When the system acts on behalf of multiple users, recalling information indefinitely and sharing it without limitations among multiple nodes is likely to generate privacy concerns, security concerns, and agency concerns (the AI system not acting in the best interest of a given user). This is not to say that no information should be remembered or shared; designing the right information structure is a nontrivial optimization problem that is closely tied to designing agent identities in multiagent systems. For example, when two nodes of the system share no information, or when a single node's memory is erased as it begins to represent a new user, we may conceive of this as resulting in multiple agents. These issues require new frameworks and solutions.

[3]Even some games of perfect recall would be problematic to let human subjects play, because the *timing* of the game would necessarily leak information (Jakobsen, Sørensen, and Conitzer 2016).

Beyond Causal Decision Theory

In light of the issues discussed so far, one may question whether traditional approaches to decision and game theory are an ideal fit for the problem of designing artificially intelligent agents. One concrete new form of decision theory that has been proposed in this context is that of *functional decision theory (FDT)* (Levinstein and Soares 2017). This is in contrast to the more standard decision theory that dominates economic theory, known in the philosophy literature as *causal decision theory (CDT)*, but also in contrast to *evidential decision theory (EDT)*, another theory that has been thoroughly studied in the philosophy literature. FDT shares some features with EDT, so let us discuss EDT first.

In EDT, when one evaluates a particular contemplated choice of action, one first updates one's beliefs, *conditioning on the fact that one has made that particular choice*. For example, suppose one is playing a game of prisoner's dilemma against someone very similar to oneself. One may reason as follows: "Given that the other reasons so similarly to me, if I choose defect, then probably so will she; whereas if I choose cooperate, then probably so will she. Hence, conditional on me cooperating, my expected utility is higher than conditional on me defecting. So I should cooperate." This is the type of reasoning that EDT endorses. A proponent of CDT, by contrast, will dispute this analysis, and point out that even if the decisions are *correlated*, one player's action has no *causal* effect on the other's; hence, defection is rational because it is better regardless of what the other does.

The vast majority of economists and game theorists subscribe (often unknowingly) to CDT. I, too, have argued against EDT, on account of a Dutch book argument in the Sleeping Beauty problem (Conitzer 2015)—a problem that will be discussed below. But one may well feel that there is something right about this EDT analysis, when we look at certain contexts involving decisions by AI agents. Consider (say) two self-driving cars of the same make, representing their owners and passing by each other. If they each have a choice between Aggressive and Defensive, where the former is a dominant strategy but both playing the latter would lead to higher utilities overall, one may feel that in some sense each car *should* play Defensive, even from a selfish perspective, because the other car, *running the same algorithm*, is sure to do the same. There is a *logical dependency* between one car's action and the other's. Taking such logical dependencies into account is arguably the fundamental idea behind FDT. FDT also generally has agents play *as they would have precommitted to play*. Now, it seems clear (under any one of these decision theories) that the *party designing the algorithm* that is used to run one or more agents has a strong form of commitment power and might as well use it. But FDT makes far stronger recommendations. For example,[4] imagine that you face an entity that can predict your actions perfectly. It will give you $100 if it believes you would otherwise take an action that is disastrous to everyone (say, the equivalent of -$1000). FDT says that you should take the disastrous action *even if you only find out that you are play-*

[4]... based on an e-mail conversation with Levinstein and Soares, following a recent talk by the former at Duke.

ing this game after the entity has already decided not to give you the money. This all-too-late extortion attempt seems to be an unacceptable implication of the theory.

Nevertheless, it seems that there may be something to this general line of thought, and perhaps an improved theory would be widely useful in the design of AI systems. For example, we already see that in real applications of game-theoretic AI, such as those in security domains (Pita et al. 2009; Yin et al. 2012; An et al. 2012; Fang et al. 2016), what is computed is an optimal mixed strategy to *commit* to (Conitzer and Sandholm 2006; von Stengel and Zamir 2010), rather than, say, a Nash equilibrium. The related idea of strategic agents reading each other's code has also already been investigated by AI researchers (Tennenholtz 2004; Oesterheld 2018), as has the idea of agents that make commitments that are conditional on commitments made by others (Kalai et al. 2010). All this results in blurry boundaries between the agents themselves, the parties who design them, and the parties that they represent. Can we discover a better theory of how to conceive of the agents and how they should make decisions? In my view, it will likely be best to build up from restricted theories in specific settings, such as those in use in security domains. This approach is likely to generate well-defined computational problems, for which efficient algorithms may be broadly useful.

Designing Agents' Beliefs

In this final section, we discuss the problem of designing agents' beliefs (or, more accurately, designing their belief formation process). This may, at first glance, not seem like a well-motivated problem: should we not just aim for our agents to hold probabilistic beliefs that reflect reality as well as possible? As it turns out, in the context of agents with imperfect recall, there is still controversy over what the right beliefs for these agents to hold are, even when the probabilistic process governing how the world evolves is known.

In particular, imperfect recall results in nontrivial problems regarding *self-locating beliefs*. The paradigmatic example is the *Sleeping Beauty* problem (Elga 2000). The Sleeping Beauty problem proceeds as follows. A (consenting!) subject in an experiment—we will call her Sleeping Beauty—is given drugs on Sunday that will make her fall asleep. Then, the experimenter tosses a coin. If the coin lands Heads, Beauty will be briefly awoken on Monday only. If it lands Tails, she will be briefly awoken on Monday, made to sleep again, and then again briefly awoken on Tuesday. The key aspect of the problem is that (due to all the drugs) Beauty cannot distinguish between the three different types of awakening—i.e., Heads/Monday, Tails/Monday, and Tails/Tuesday. Specifically, she is unable to keep track of time, and will have forgotten her Monday awakening if and when she is awoken on Tuesday. The problem is as follows. Imagine you are Beauty and you have just been awoken in this experiment. What is your subjective probability that the coin came up Heads?[5]

This problem splits people into two primary camps. One believes that the answer should be $1/2$ (the "halfers"); the other, $1/3$ (the "thirders"). One argument for the halfer position is that Beauty always knew she would be awoken at *some* point; thus, she has received no new evidence since Sunday, at which point she clearly should have believed there was a probability of $1/2$ that the coin would come up Heads. One argument for the thirder position is a frequentist one: if this experiment is repeated every week, then in the long run the fraction of Heads awakenings approaches $1/3$. Many other arguments have been given on both sides. One family of arguments relies on investigating *decisions* resulting from these beliefs. Hitchcock (2004) argues that halfers are vulnerable to a Dutch book—a combination of bets that they would all accept, but that together result in a sure loss. However, Draper and Pust (2008) point out that this is only the case for halfers that adopt CDT; halfers that adopt EDT do not fall for the Dutch book. Building on this, Briggs (2010) proves that, in a class of settings, thirders who accept CDT and halfers who accept EDT are both immune to Dutch books. However, if we extend to a slightly broader class of settings, anyone doing what EDT would seem to recommend is vulnerable to a Dutch book (Conitzer 2015).

Nevertheless, Briggs' result suggests an interesting possibility. We may not be able to definitively resolve the Sleeping Beauty puzzle without certain assumptions/insight about the metaphysics of time and self, due to the unusual nature of Beauty's evidence. (When determining my posterior beliefs, how should I condition on the fact that "*I* have *now* just been awoken"?[6]) This is worth pondering, but for the practical purpose of designing AI systems that display some types of imperfect recall—for example, because nodes do not share information, for privacy purposes—perhaps it does not matter what theory of self-locating belief we use, as long as we combine it with the right decision theory.

One may go one step further and argue that perhaps we should not bother with probabilistic beliefs at all, and rather just analyze directly which policies are most effective, perhaps somewhat along the lines of FDT as discussed earlier. But in sufficiently complex environments, it sometimes helps enormously to be able to separate out belief formation and the process of making decisions based on those beliefs. This is true in part because many tools have been developed for both probabilistic reasoning and decision making in the AI literature. It seems that these tools should generalize to problems that involve self-locating belief. Achieving this would require appropriately extending both standard representation schemes for probabilistic uncertainty (e.g., Bayesian networks) and algorithms for reasoning with them.

[5] For those not inherently motivated by philosophical conundrums, consider the following reinterpretation of the problem: a car has low-level autonomy with AI that is called upon to intervene only in certain dangerous situations, but it does not keep records of these events. We know that half the cars have careful drivers who will get into such a situation once ("Heads"), and the other half have careless drivers who will get into such a sitation twice ("Tails"). When called upon to act, what should the AI conclude is the probability that the driver is careful (which may well be very relevant to the driving decisions it is about to take)?

[6] Recent literature in philosophy addresses the apparent datum that a single subjective experience is *present* (Valberg 2007; Hare 2007; 2009; 2010; Hellie 2013; Merlo 2016; Conitzer 2018).

Acknowledgments

I am thankful for support from NSF under awards IIS-1814056 and IIS-1527434.

References

Abel, D.; MacGlashan, J.; and Littman, M. L. 2016. Reinforcement learning as a framework for ethical decision making. In *AAAI Workshop on AI, Society, & Ethics*.

An, B.; Shieh, E.; Tambe, M.; Yang, R.; Baldwin, C.; DiRenzo, J.; Maule, B.; and Meyer, G. 2012. PROTECT - A deployed game theoretic system for strategic security allocation for the United States Coast Guard. *AI Magazine* 33(4):96–110.

Bonnefon, J.-F.; Shariff, A.; and Rahwan, I. 2016. The social dilemma of autonomous vehicles. *Science* 352(6293):1573–1576.

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Briggs, R. 2010. Putting a value on Beauty. In Tamar Szabó Gendler and John Hawthorne, editors, *Oxford Studies in Epistemology: Volume 3*. 3–34.

Conitzer, V., and Sandholm, T. 2006. Computing the optimal strategy to commit to. *ACM EC*, 82–90.

Conitzer, V.; Freeman, R.; Brill, M.; and Li, Y. 2016. Rules for choosing societal tradeoffs. *AAAI*, 460–467.

Conitzer, V.; Sinnott-Armstrong, W.; Borg, J. S.; Deng, Y.; and Kramer, M. 2017. Moral decision making frameworks for artificial intelligence. *AAAI*, 4831–4835.

Conitzer, V.; Brill, M.; and Freeman, R. 2015. Crowdsourcing societal tradeoffs. *AAMAS*, 1213–1217.

Conitzer, V. 2015. A Dutch book against sleeping beauties who are evidential decision theorists. *Synthese* 192(9):2887–2899. DOI 10.1007/s11229-015-0691-7.

Conitzer, V. 2018. A Puzzle about Further Facts. *Erkenntnis*. https://doi.org/10.1007/s10670-018-9979-6.

Draper, K., and Pust, J. 2008. Diachronic Dutch Books and Sleeping Beauty. *Synthese* 164(2):281–287.

Elga, A. 2000. Self-locating belief and the Sleeping Beauty problem. *Analysis* 60(2):143–147.

Fang, F.; Nguyen, T. H.; Pickles, R.; Lam, W. Y.; Clements, G. R.; An, B.; Singh, A.; Tambe, M.; and Lemieux, A. 2016. Deploying PAWS: Field optimization of the protection assistant for wildlife security. *IAAI*.

Freedman, R.; Borg, J. S.; Sinnott-Armstrong, W.; Dickerson, J. P.; and Conitzer, V. 2018. Adapting a kidney exchange algorithm to align with human values. *AAAI*.

Fudenberg, D., and Levine, D. 1998. *The Theory of Learning in Games*. MIT Press.

Greene, J.; Rossi, F.; Tasioulas, J.; Venable, K. B.; and Williams, B. C. 2016. Embedding ethical principles in collective decision support systems. *AAAI*, 4147–4151.

Hadfield-Menell, D.; Dragan, A. D.; Abbeel, P.; and Russell, S. J. 2017. The off-switch game. *IJCAI*, 220–227.

Hare, C. 2007. Self-Bias, Time-Bias, and the Metaphysics of Self and Time. *Journal of Philosophy* 104(7):350–373.

Hare, C. 2009. *On Myself, And Other, Less Important Subjects*. Princeton University Press.

Hare, C. 2010. Realism About Tense and Perspective. *Philosophy Compass* 5(9):760–769.

Hellie, B. 2013. Against Egalitarianism. *Analysis* 73(2):304–320.

Hitchcock, C. 2004. Beauty and the bets. *Synthese* 139(3):405–420.

Jakobsen, S. K.; Sørensen, T. B.; and Conitzer, V. 2016. Timeability of extensive-form games. *ITCS*, 191–199.

Kalai, A. T.; Kalai, E.; Lehrer, E.; and Samet, D. 2010. A commitment folk theorem. *Games and Economic Behavior* 69(1):127–137.

Kroer, C., and Sandholm, T. 2016. Imperfect-recall abstractions with bounds in games. *ACM EC*, 459–476.

Lanctot, M.; Gibson, R. G.; Burch, N.; and Bowling, M. 2012. No-regret learning in extensive-form games with imperfect recall. *ICML*.

Levinstein, B., and Soares, N. 2017. Cheating Death in Damascus.

Merlo, G. 2016. Subjectivism and the Mental. *Dialectica* 70(3):311–342.

Noothigattu, R.; Gaikwad, S. N. S.; Awad, E.; D'Souza, S.; Rahwan, I.; Ravikumar, P.; and Procaccia, A. D. 2018. A voting-based system for ethical decision making. *AAAI*.

Oesterheld, C. 2018. Robust program equilibrium. *Theory and Decision*. DOI 10.1007/s11238-018-9679-3.

Pita, J.; Jain, M.; Ordóñez, F.; Portway, C.; Tambe, M.; and Western, C. 2009. Using game theory for Los Angeles airport security. *AI Magazine* 30(1):43–57.

Tegmark, M. 2017. *Life 3.0: Being Human in the Age of Artificial Intelligence*. Knopf.

Tennenholtz, M. 2004. Program equilibrium. *Games and Economic Behavior* 49:363–373.

Valberg, J. J. 2007. *Dream, Death, and the Self*. Princeton University Press.

von Stengel, B., and Zamir, S. 2010. Leadership games with convex strategy sets. *Games and Economic Behavior* 69:446–457.

Waugh, K.; Zinkevich, M.; Johanson, M.; Kan, M.; Schnizlein, D.; and Bowling, M. H. 2009. A practical use of imperfect recall. *SARA*.

Yin, Z.; Jiang, A. X.; Tambe, M.; Kiekintveld, C.; Leyton-Brown, K.; Sandholm, T.; and Sullivan, J. P. 2012. TRUSTS: Scheduling randomized patrols for fare inspection in transit systems using game theory. *AI Magazine* 33(4):59–72.

Zhang, H., and Conitzer, V. 2019. A PAC framework for aggregating agents' judgments. *AAAI*.