Incomplete Contracting and AI Alignment

Dylan Hadfield-Menell
University of California, Berkeley
Center for Human-Compatible AI
dhm@eecs.berkeley.edu

Gillian K. Hadfield
University of Toronto
Vector Institute for AI; OpenAI
Center for Human-Compatible AI
g.hadfield@utoronto.ca

ABSTRACT

We suggest that the analysis of incomplete contracting developed by law and economics researchers can provide a useful framework for understanding the AI alignment problem and help to generate a systematic approach to finding solutions. We first provide an overview of the incomplete contracting literature and explore parallels between this work and the problem of AI alignment. As we emphasize, misalignment between principal and agent is a core focus of economic analysis. We highlight some technical results from the economics literature on incomplete contracts that may provide insights for AI alignment researchers. Our core contribution, however, is to bring to bear an insight that economists have been urged to absorb from legal scholars and other behavioral scientists: the fact that human contracting is supported by substantial amounts of external structure, such as generally available institutions (culture, law) that can supply implied terms to fill the gaps in incomplete contracts. We propose a research agenda for AI alignment work that focuses on the problem of how to build AI that can replicate the human cognitive processes that connect individual incomplete contracts with this supporting external structure.

CCS CONCEPTS

• **Applied computing** → **Economics**; *Law*; • **Computing methodologies** → *Cooperation and coordination*; *Inverse reinforcement learning*.

KEYWORDS

value alignment; incomplete contracting; principal-agent problems

ACM Reference Format:
Dylan Hadfield-Menell and Gillian K. Hadfield. 2019. Incomplete Contracting and AI Alignment. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '19), January 27–28, 2019, Honolulu, HI, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3306618.3314250

1 INTRODUCTION

When we design and deploy an AI agent, we specify what we want it to do. In reinforcement learning, for example, we specify a reward function, which tells the agent the value of all state and action combinations. Good algorithms then generate AI behavior that performs well according to this reward function. The AI alignment problem arises because of differences between the specified reward function and what relevant humans (the designer, the user, others affected by the agent's behavior) actually value. AI researchers intend for their reward functions to give the correct rewards in all states of the world so as to achieve the objectives of relevant humans. But often AI reward functions are—unintentionally and unavoidably—misspecified. They may accurately reflect human rewards in the circumstances that the designer thought about but fail to accurately specify how humans value all state and action combinations.

AI alignment has a clear analogue in the human principal-agent problem long studied by economists and legal scholars. In these settings a human agent is supposed to take actions that achieve a principal's objectives. The ideal way to align principal and agent is to design a *complete contingent contract* [51]. This is an enforceable agreement that specifies the reward received by the agent for all actions and states of the world. The contract could be enforced by monetary transfers or punishments imposed by a coercive institution, such as a court. Or it could be enforced by a private actor or group of actors who penalize contract violations by, for example, imposing social sanctions or cutting off valuable relationships; the latter contracts are known as *relational contracts* [13, 29, 31, 32]. A complete contingent contract implements desired behavior by the agent in all states of the world. It perfectly aligns the agent's incentives with the principal's objective.

In this paper, [1]. we suggest that the analysis of incomplete contracting developed by law and economics researchers can provide a useful framework for understanding the AI alignment problem and help to generate a systematic approach to finding solutions. We briefly highlight some technical results from the economics literature on incomplete contracts that may provide insights for AI alignment researchers. Our core contribution, however, is to bring to bear an insight that economists have been urged to absorb from legal scholars and other behavioral scientists: the fact that human contracting is supported by substantial amounts of external structure, such as generally available institutions (culture, law) that can supply implied terms to fill the gaps in incomplete contracts. We propose a research agenda for AI alignment work that focuses on the problem of how to build AI that can replicate the human cognitive processes that connect individual incomplete contracts with this supporting external structure.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
AIES '19, January 27–28, 2019, Honolulu, HI, USA
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-6324-2/19/01…$15.00
https://doi.org/10.1145/3306618.3314250

[1]This paper is a condensed version of a more fully elaborated analysis of the parallels between incomplete contracting and the AI alignment problem, available at https://arxiv.org/abs/1804.04268

2 THE FUNDAMENTAL PROBLEM OF MISALIGNMENT

Economies are built on specialization and the division of labor–meaning that the production and allocation of different things of value needs to be coordinated across a group of humans. The challenge of aligning the interests of one actor with others' is at the core of modern economic theory. The first of two fundamental welfare theorems [5, 11] states that if markets are complete (all possible trades can be made including those that involve future goods and services and third-party effects) and perfectly competitive (there are no transaction costs, all information is common knowledge, and no one holds market power), then voluntary trading in a market economy will produce a result that is Pareto efficient: there is no way to reallocate resources or goods so as to make someone better off without making someone else worse off. This is a form of alignment: the market outcome aligns with the solution that maximizes a social welfare function that aggregates the values of all members of the economy, weighted by their initial endowments of goods.[2] The second welfare theorem then states that for any distributive goal–for any final distribution of goods or social welfare weights that society chooses–there is an initial allocation of endowments (including labor) such that a perfectly competitive and complete market economy will produce that final distribution. In theory, perfectly competitive and complete markets serve as a mechanism to align individual decisions about production and trade so as to maximize a social welfare function.

Despite the centrality of these core welfare theorems, most work in economics focuses on the failure of markets to be perfectly competitive and complete. The welfare theorems for perfect markets merely provide a framework for thinking about how to design production and allocation systems–markets, organizations, laws–to achieve better outcomes from a social welfare point of view. Departures from perfect and complete markets introduce costs due to distortion, that is, a failure of alignment. Some things that humans value cannot be fully traded. Most fundamentally, there is no coherent social welfare function that is based exclusively on subjective assessments of own utility; any coherent social welfare function requires collective judgments to be made about what values to pursue [5, 41] and so there is inevitable "misalignment" with the values of some humans.

Misalignment in the human principal agent setting is responsible for the economic loss associated with delegation of decisions over productive effort. In a perfect frictionless world, where all factors that affect outcomes are common knowledge and any agreement between actors can be costlessly written and enforced, voluntary agreements in which an agent agrees to take certain actions and a principal agrees to compensate the agent in particular ways will align the interests of agent and principal. Core results in the theory of contracts then explore whether it is possible to align interests (achieve the "first-best" promised by the fundamental welfare theorems) when there is hidden information such as when an agent has private information about the cost of taking an action (adverse selection) or about the action chosen (moral hazard) [28]. Sometimes this is indeed possible. But in general, the first-best is not achievable: the optimal contract trades off giving the agent incentives to be productive (the size of the pie) and the achievement of the principal's goal or utility (share of the pie.)

Misalignment in the design of artificially intelligent agents can be thought of in parallel terms. AI designers, like contract designers, are faced with the challenge of achieving intended goals in light of the limitations that arise from translating those goals into implementable structures to guide agent behavior (learning algorithms and reward functions). An AI is misaligned whenever it chooses behaviors based on a reward function that is different from the true welfare of relevant humans. We see misalignment as the general description of a wide variety of problems that go by different names in AI research. [3] collect a set of cases that they refer to as "accidents": situations in which a human designer has an objective in mind but the system as designed and deployed produces "harmful and unexpected results." They propose several mechanisms producing such accidents: negative side-effects, reward hacking, limited capacity for human oversight, differences between training and deployment environments, and uncontrolled or unexpected exploration after deployment. Alignment problems also arise because of the difficulty of representing and implementing human values. The problems of fairness and bias in machine learning algorithms are fundamentally alignment problems. The technical literature here (see, e.g. [21, 27, 30, 52, 53]) seeks to develop techniques to align algorithmic decisions with complex human goals such as discriminating between prospective employees on ability but not gender or race. AI safety problems such as safe interruptibility [37], the off-switch game [18] and corrigibility [47] are also alignment problems: these are efforts to ensure that AI agents value shut down or modification of their reward functions in the same way that humans do or at least are indifferent to such efforts. And at the most general level, the question of how to elicit and aggregate preferences when there are multiple humans affected by the behavior of an artificial agent [38] is an alignment problem. Indeed, it is the basic alignment problem addressed by the fundamental theorems of welfare economics.

3 REASONS FOR MISALIGNMENT

It is natural to think that misalignment between agent and principal is just an error in design. And indeed, sometimes misalignment in the human principal agent setting is the result of bad contract drafting and sometimes in the context of artificial intelligence it is the result of straightforward misspecification of what the designer wants. But these are not the particularly interesting or challenging cases of misalignment. In this section we briefly collect the reasons for contract incompleteness and then provide what we think are the parallel reasons for reward misspecification in AI.

The most commonly cited reason that contracts are incomplete is because completeness is practically impossible or costly: contract designers might fail to identify all circumstances that affect the value of the contract[45, 51], choose not to invest in the costly cognitive effort of discovering, evaluating and drafting contract terms to cover all circumstances [40, 42, 43], leave out terms that are costly to enforce [20, 25, 39], or leave out contingencies and/or actions because they cannot be verified by enforcers at reasonable cost [16, 35]. These reasons for incompleteness seem to us reasonably to translate over to the AI context. Rewards may fail to address

[2] The weighting is due to the use of the Pareto criterion.

all relevant circumstances because designers simply did not (and perhaps could not) think of everything (see, e.g., [19], they may have deliberately chosen not to invest additional effort to identify or code for possible contingencies, some reward structures are, given the state of the art, simply not implementable. If we think of a contract as an implemented reward structure for a human agent, the analog to non-contractibility is a learning problem that is not solvable with known techniques.

In some settings, contractual completeness is feasible but not optimal. These are cases in which information at the time of contracting is incomplete and new information is anticipated in the future. Contract designers might choose to write an incomplete contract which they expect to renegotiate in the future once more information becomes available [9, 20] or to be filled in by a third-party adjudicator with better information in the future[17, 44]. This is analogous to the case in which an AI designer has to choose between developing a more complete reward structure today and deferring decisions about how to build rewards until more information has been learned. The problem labeled "safe exploration" by [3] seems to fit this description.

Finally, economists, and many legal scholars, have also proposed that contracts may be incomplete because of strategic behavior: a party with private information about a missing contingency may not prompt contracting to cover the contingency because doing so will reveal private information that reduces the value of the contract [6, 49], parties may choose not to cover all contingencies because learning about them would be biased and wasteful [50], or parties may choose not to include all known and contractible contingencies in order to control strategic behavior in response to other noncontractible contingencies [8]. Strategic considerations on the part of human AI designers could also lead them to choose to develop agents that are deliberately not given a complete specification of everything the designer cares about. This type of technique is common in the domain adaptation literature, for example, which tackles the problem of what to do when you have a small amount of data from the setting/distribution that is of interest but can obtain a lot of data for training purposes from a different setting/distribution [12]. [3] call this "adversarial blinding". Strategic incompleteness in reward design may also become relevant in more advanced systems than those we have today [4] if we contemplate sequential reward design with a powerful agent. If a robot predicts that the human may rewrite the reward structure, for example, then the robot, currently implementing the initial reward, may behave strategically–withholding information–to influence the rewriting so as to preserve the initial reward structure.

Table 1 summarizes the parallels between the reasons for contractual incompleteness and the reasons for reward misspecification.

4 INSIGHTS FOR AI ALIGNMENT FROM THE ECONOMICS OF INCOMPLETE CONTRACTING

In this section, we provide a brief overview of key results in the economic theory of incomplete contracting to identify potential insights for AI researchers.[3]

[3]More detailed analysis is available at https://arxiv.org/abs/1804.04268

The potential for AIs to strategize in order to achieve goals as embodied in their initial design has been a focus of the study of superintelligence [10, 36, 47], envisioning the potential for what we will call *strongly strategic* AI systems to rewrite their reward functions, alter their hardware, or manipulate humans. But the value of a strategic formulation of AI behavior does not arise only in these futuristic settings. It arises in any routine setting in which there is a divergence between the AI's stated reward function and true or intended human value [18, 19]. With this divergence in rewards, the AI and the human are inherently engaged in the strategic game of each trying to take actions that allow them to do well against their own reward function, even if that reduces value for the other. We will call these systems *weakly strategic*.

4.1 Weakly Strategic AI

We begin by highlighting two key results from the economic literature that we think can contribute to problems in existing, weakly strategic, AI systems.

4.1.1 Property Rights. A core result in the early incomplete contracting literature is that when complete contingent contracts are not possible, the joint value produced by two economic actors may be maximized through the allocation of property rights over productive assets [16, 23]. From an economic point of view, the allocation of property rights is just a means of determining a reward function. Granting "property rights" to an AI, then, could be understood as imbuing an AI with the reward function associated with the ultimate rewards generated by a productive activity in which it participates. Seen in this light, the insight from the analysis of property rights and incomplete contracts is one that is already at the heart of the AI alignment challenge.

4.1.2 Measurement and Multi-tasking. An important reason that contracts are incomplete is the difficulty of specifying how an action is to be measured or conditioning payoffs on a particular measurement. A key problem arises when an agent engages in multiple tasks, and the tasks are differentially measurable or contractible. [24] and [7] show that it may be better to reduce the quality of incentives on a measurable task below what is feasible, in order to reduce the distortion introduced in an unmeasurable task. More generally, sometimes it is better for a contract not to include easily contractible actions in order not to further distort incentives with respect to non-contractible actions. The lesson for AI is that a singular focus on improving performance on the measurable task may degrade performance on the unmeasurable.

4.2 Strongly Strategic AI

The implications of the economic analysis of incomplete contracting for strongly strategic AI are more speculative, because we don't know how (or if) such systems will evolve. But we set out briefly lines of research that may be of interest to researchers thinking through the challenge of managing powerful AI that can act in an overtly strategic way to resist changes to its reward function, for example, or evade shut-down.

4.2.1 Control Rights. Property rights over assets generalize to decision or control rights: any authority to make decisions about actions in circumstances not controlled by an enforceable contract.

Why are contracts incomplete?	Why are rewards misspecified?
Bounded rationality (can't think of all contingencies)	Bounded rationality (negative side effects)
Costly cognition/drafting	Costly engineering/design
Non-contractibility (variables not describable/verifiable to enforcer)	Non-implementability (unsolved learning problems)
Planned renegotiation	Planned iteration on rewards
Planned completion by third party in event of dispute	Planned completion by third party
Strategic behavior	Adversarial blinding, reward preservation

Table 1: Parallel reasons for incompleteness and misspecification

[1] present a model in which an agent has reduced incentives to acquire information about the environment if the principal has formal authority (the right to make final decisions) *and the information needed to exercise it.* Paradoxically, the principal might be able to improve the incentives of the agent to acquire knowledge by remaining uninformed and thereby making a credible commitment not to intervene. This result has possible implications for a strongly strategic AI's incentives to share information with human controllers. One interpretation of [18] is as a model not only of the incentive of a robot to disable its off switch but also of its incentive to share with a human the information needed to exercise off-switch authority. System designers may need to take into account the tradeoffs between generating information necessary for meaningful human oversight and minimizing the impact of human information on the robot's performance.

4.2.2 Costly Signaling. The economic literature on costly signaling looks at the ways in which contracts can be designed so as to elicit private information from agents. The work originates with a seminal paper by [48] which shows that if an employer offers a high wage contract to those who meet or exceed a specified educational level and a lower wage contract to those who do not, prospective job applicants can be induced to sort themselves such that high ability workers accept the high wage contract and low ability workers accept the low wage contract, provided that education is sufficiently more costly for low ability than high ability applicants. This line of analysis may have an interesting application to AI alignment. A key information problem for a human designer is to know a robot's current estimate of the reward function, which can be more or less aligned with the human's true utility. If intervention substituting the human's preferred action for the robot's will be more costly for the less aligned robot than the more aligned one, it might be possible to design systems in which the willingness of a strongly strategic AI to seek human input serves as a signal of the AI's alignment.

4.2.3 Renegotiation. There is a basic trade-off, reproduced in the AI setting, between specifying behaviors for an agent *ex ante* with incomplete information and specifying optimal behaviors *ex post* once more information about the state of the world is available. This creates a risk of hold-up. Even with a contract, if it is ultimately discovered that the action called for is not optimal, there is an incentive to renegotiate. In the standard incomplete contracting setting, this means having to pay the agent to agree to shift from the old contract to a new one. A key insight is that the provisions of the initial contract set the terms on which the new contract is bargained [22]. This suggests that strongly strategic AI systems may need to be effectively "bought out:" incentivized to shift from a reward function they were originally given to one that a human later discovers is closer to the truth.

5 INSIGHTS FOR AI ALIGNMENT FROM THE LAW OF INCOMPLETE CONTRACTING

Contracts do not exist in a vacuum; they come heavily embedded in social and institutions structures [15]. At a minimum, they depend on shared language and organized structures for enforcement: formal enforcement through courts and coercive authorities and informal enforcement through social sanctions such as collective (coordinated) criticism and exclusion from valuable relationships. Incomplete contracts depend even more extensively on these external, third-party institutions: not only to enforce contractual terms but to supply contractual terms by interpreting ambiguous terms and filling in gaps. This important point has been emphasized by legal scholars, in a field known as relational contracting, for several decades [14, 31–34]. [2, 26, 51] were the first economic treatments to focus on this aspect of the legal analysis of relational contracts.

The lesson that [15] pressed on economists in the early stages of the analysis of incomplete contracting, however, seems equally apt for AI researchers tackling the problem of alignment today: alignment problems cannot be solved without support from external normative structure.

Consider a simple example posed by [3]: a robot learns to move boxes from one side of a room to another in a training environment that lacks obstacles. (See Figure 1.) When deployed, a vase appears in the path the robot has learned to use. The robot that is rewarded for transferring boxes and not penalized for knocking over the vase will ignore the vase. [3] use this as an example of negative side-effects: unintended consequences arising from a failure to include relevant features in the robot's reward.

Consider now what happens if we hire a human agent to carry boxes. Suppose the contract is a direct analogue of the robot's reward function. It says that the agent will be paid a certain amount for every box carried to the other side of the room. It says nothing about knocking over vases, and there are no vases about when the agent is hired. What happens when a vase appears? Easy: the agent will walk around the vase. Why?

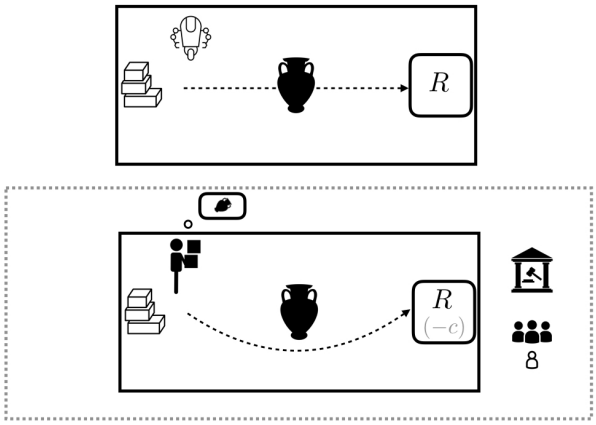

Figure 1: When a robot is given a reward function that specifies a reward only based on moving boxes, it will ignore a vase that appears in the path [3]. If a human agent is given a contract that pays only for moving boxes, she will interpret her contract to include an implied term that penalizes knocking over the vase.

The reason is that the incomplete agreement that says nothing about vases is not the entire contract between principal and agent. Human contracts are not limited to the express terms; they also include implied terms. In particular, in these circumstances, the contract implicitly contains a term that imposes a cost on breaking the vase. If the agent ignores the vase, the reward the agent receives will be reduced by some amount: the agent might be charged for the breakage, she may suffer future income losses as a result of being fired or earning a bad reputation in the labor market, she may suffer psychic pain as a result of being criticized, or she may suffer the discomfort of feeling guilty, ashamed, or incompetent for having done something wrong.

These implied terms represent a type of "common sense" reasoning — reasoning about the extensive normative structure in which human contracting is embedded. Our human contract arises in an environment filled with rules and institutions that resolve, precisely, questions such as: was it wrong for the agent to knock over the vase while carrying out this task? Some of these rules might come in the form of cultural norms–classifications that arise as emergent features from repeated interaction and discussion among participants in a group. Others of these rules are the product of formal dispute-resolution systems of law and adjudication that humans administer to fill in the gaps in incomplete contracts.

We conjecture that any robust solution to the AI alignment problem will also require the recruitment of normative resources external to the reward structure designed for any particular application.

We do not mean by this embedding into the AI the particular *norms and values* of a human community. We think this is as impossible a task as writing a complete contract. Human norms and values are highly variable and deeply contextual. Suppose, for example, that the contract puts a time limit on the movement of the boxes and the agent can't move them all in time without knocking over the vase. Or the boxes contain medical equipment that is urgently needed to treat trauma patients and nobody cares about broken vases. Or the vase is a sacred object and the people who need the medical treatment also believe that knocking over the vase will anger the gods and only make recovery less likely. There is no easy answer to the question of whether it is okay or not okay to break the vase.

We usually refer to adapting to these different contexts as "common sense," but it is important to emphasize that this is common sense *about what actions society will sanction.* The human agent's capacity to infer implied terms about the values associated with breaking the vase is a product of the human's ability to interact with and participate in normative social structure. Humans are endowed with the cognitive architecture needed to read and predict the responses of this normative structure. The human agent avoids the vase in those circumstances in which she concludes that the community will sanction breaking it and plows on through otherwise.

Aligning AI with human values, then, will require figuring out how to build the technical tools that will allow a robot to replicate the human agent's ability to read and predict the responses of human normative structure, whatever its content.

Building AI that can supplement their designed rewards with implied rewards from community normative structure also will require building tools that allow a robot to assign negative weight to actions classified as sanctionable by the community. Human agents assign costs to taking actions that violate implied terms in contracts for many reasons. Some of these penalties are formally imposed by enforcement institutions such as courts (contract damages, for example). Others are imposed informally through coordinated community action: refusing to hire or do business or engage socially with someone who is judged by formal or informal standards to have breached a contract, for example. These externally-imposed penalties seem difficult to transpose to the AI context.

But humans also internalize social penalties. An important part of human development is the building of the cognitive architecture for experiencing negative emotions such as distress, shame, and guilt in response to a real or imagined public judgment of rule violation. Adam [46] famously referred to this as the capability to view one's own conduct as if through the eyes of an "impartial spectator." This cognitive architecture–creating the mental buttons that external normative criticism can press to change behavior, so to speak–would seem to have a natural analog in the design of an artificial intelligence–assigning loss to conditions in which the prediction is made that an action, in context, would be judged by external human communities to be wrongful.

6 CONCLUSION

The alignment of artificially intelligent agents with human goals and values is a fundamental challenge in AI research. It is also the fundamental challenge of organizing human economic interaction in an economy built on specialization and the division of labor–in which humans are tasked with taking actions that generate costs and benefits for other humans. By recognizing and elaborating the parallels between the challenge of incomplete contracting in the

human principal-agent setting and the challenge of misspecification in robot reward functions, this paper provides AI researchers with a different framework for the alignment problem. That framework urges researchers to see reward misspecification as fundamental and not merely the result of poor engineering. Doing so, as we show, generates insights both for the analysis of current, weakly strategic, AI systems and potential, strongly strategic, systems. Our most important claim is that aligning robots with humans will inevitably require building the technical tools to allow AI to do what human agents do naturally: import into their assessment of rewards the costs associated with taking actions tagged as wrongful by human communities. These are the lessons learned by economists and legal scholars over the past several decades in the context of incomplete contracting. They are lessons available also to AI researchers.

REFERENCES

[1] Philippe Aghion and Jean Tirole. 1997. Formal and Real Authority in Organizations. *Journal of Political Economy* 105 (1997), 1–29.

[2] Armen A. Alchian and Harold Demsetz. 1972. Production, Information Costs, and Economic Organization. *The American Economic Review* 62, 5 (1972), 777–795.

[3] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. *arXiv preprint arXiv:1606.06565* (2016).

[4] Stuart Armstrong. 2015. Motivated Value Selection for Artificial Agents. In *AAAI Workshop: AI and Ethics*.

[5] Kenneth J. Arrow. 1951. An Extension of the Basic Theorems of Classical Welfare Economics. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*.

[6] Ian Ayres and Robert Gertner. 1989. Filling Gaps in Incomplete Contracts: An Economic Theory of Default Rules. *The Yale Law Journal* 99 (1989), 87–130. Issue 1.

[7] George Baker, Robert Gibbons, and Kevin J. Murphy. 1994. Subjective Performance Measures in Optimal Incentive Contracts. *The Quarterly Journal of Economics* 109, 4 (1994), 1125–1156.

[8] B. Douglas Bernheim and Michael D. Whinston. 1998. Incomplete Contracts and Strategic Ambiguity. *The American Economic Review* 88, 4 (1998), 902–932.

[9] Antoine Bolton, Patrick anhd Faure-Grimaud. 2010. Satisficing Contracts. *The Review of Economic Studies* 77 (2010), 937–971.

[10] Nick Bostrom. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford.

[11] Gerard Debreu. 1959. *A Theory of Value: An Axiomatic Analysis of Economic Equilibrium*. John Wiley, New York.

[12] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-Adversarial Training of Neural Networks. *The Journal of Machine Learning Research* 17, 1 (2016), 2096–2030.

[13] Ricard Gil and Giorgio Zanarone. 2017. Formal and Informal Contracting: Theory and Evidence. *Annual Review of Law and Social Science* 13 (2017), 141–159.

[14] Charles J. Goetz and Robert E. Scott. 1981. Principles of Relational Contracts. *Virginia Law Review* 67, 6 (1981), 1089–1150.

[15] Mark Granovetter. 1985. Economic Action and Social Structure: The Problem of Embeddedness. *Amer. J. Sociology* 91, 3 (1985), 481–510.

[16] Sanford Jay Grossman and Oliver D. Hart. 1986. The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration. *Journal of Political Economy* 94, 4 (1986), 691–719.

[17] Gillian K. Hadfield. 1994. Judicial Competence and the Interpretation of Incomplete Contracts. *Journal of Legal Studies* 23 (1994), 159–184.

[18] Dylan Hadfield-Menell, Anca D. Dragan, Pieter Abbeel, and Stuart Russell. 2017. The Off-Switch Game. In *International Joint Conference on Artificial Intelligence*.

[19] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca. Dragan. 2017. Inverse Reward Design. In *Advances in Neural Information Processing Systems*. https://arxiv.org/pdf/1711.02827.pdf

[20] Maija Halonen-Akatwijuka and Oliver D. Hart. 2013. More is Less: Why Partiews May Deliberately Write Incomplete Contracts. http://www.nber.org/papers/w19001

[21] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*. 3315–3323.

[22] Oliver D. Hart. 1988. Incomplete Contracts and The Theory of the Firm. *Journal of Law, Economics & Organization* 4, 1 (1988), 119–139.

[23] Oliver D. Hart and John Moore. 1988. Incomplete Contracts and Renegotiation. *Econometrica* 56, 4 (1988), 755–785.

[24] Bengt Holmstrom and Paul Milgrom. 1991. Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics, & Organization* 7 (1991), 24–52.

[25] Benjamin Klein. 1980. Transaction Cost Determinants of 'Unfair' Contractual Arrangements. *The American Economic Review* 70 (1980), 356–362. Issue 2.

[26] Benjamin Klein, Robert G Crawford, and Armen A. Alchian. 1978. Vertical Integration, Appropriable Rents, and the Competitive Contracting Process. *Journal of Law and Economics* 21, 2 (1978), 297–326.

[27] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-offs in the Fair Determination of Risk Scores. In *Innovations in Theoretical Computer Science (ITCS)*.

[28] Jean-Jacques Laffont. 1989. *The Economics of Uncertainty and Information*. The MIT Press, Cambridge: Massachusetts.

[29] Jonathan Levin. 2003. Relational Incentive Contracts. *The American Economic Review* 93, 3 (2003), 835–857.

[30] Kristian Lum and James E. Johndrow. 2016. A Statistical Framework for Fair Predictive Algorithms. https://arxiv.org/abs/1610.08077

[31] Stewart Macaulay. 1963. Non-Contractual Relations in Business: A Preliminary Study. *American Sociological Review* 28, 1 (1963), 55–67.

[32] Ian R. Macneil. 1974. The Many Futures of Contracts. *Southern California Law Review* 1973-1974 (1974), 691–816.

[33] Ian R. Macneil. 1978. Contracts: Adjustment of Long-term Economic Relations Under Classical, Neoclassical, and Relational Contract Law. *Northwestern University Law Review* 72 (1978), 854–905.

[34] Ian R. Macneil. 1983. Values in Contract: Internal and External. *Northwestern University Law Review* 1983-1984 (1983), 340–418.

[35] Eric Maskin and Jean Tirole. 1999. Unforeseen Contingencies and Incomplete Contracts. *The Review of Economic Studies* 66, 1 (1999), 83–114.

[36] Stephen M. Omohundro. 2008. The Basic AI Drives. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.393.8356&rep=rep1&type=pdf

[37] Laurent Orseau and Stuart Armstrong. 2016. Safely Interruptible Agents. In *Proceedings of Uncertainty in Artificial Intelligence*. https://intelligence.org/files/Interruptibility.pdf

[38] Francesca Rossi, Kristen Brent Venable, and Toby Walsh. 2011. *A Short Introduction to Preferences Between Artificial Intelligence and Social Choice*. Morgan & Claypool Publishers, San Rafael, California.

[39] Alan Schwartz and Robert E. Scott. 2003. Contract Theory and the Limits of Contract Law. *The Yale Law Review* 113 (2003), 541–619.

[40] Robert E. Scott and George Triantis. 2006. Anticipating Litigation in Contract Design. *Yale Law Journal* 115 (2006), 814.

[41] Amartya Sen. 1985. Social Choice and Justice: A Review Article. *Journal of Economic Literature* 23 (1985), 1764–1776. Issue 4.

[42] Steven Shavell. 1980. Damage Measures for Breach of Contract. *The Bell Journal of Economics* 11, 2 (1980), 466–490.

[43] Steven Shavell. 2006. On the Writing and Interpretation of Contracts. *Journal of Law, Economics and Organization* 22 (2006), 289–311.

[44] Steven Shavell. 2006. On the Writing and Interpretation of Contracts. *Journal of Law, Economics and Organization* 22 (2006), 289–311.

[45] Herbert A Simon. 1955. A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics* 69, 1 (1955), 99–118.

[46] Adam Smith. 1759. *The Theory of Moral Sentiments*. A Millar, London.

[47] Nate Soares, Benja Fallenstein, Stuart Armstrong, and Eliezer Yudkowsky. 2015. Corrigibility. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[48] Michael Spence. 1973. Job Market Signaling. *The Quarterly Journal of Economics* 87 (1973), 355–374. Issue 3.

[49] Kathryn E. Spier. 1992. Incomplete Contracts and Signalling. *The RAND Journal of Economics* 23 (1992), 432–443. Issue 3.

[50] Jean Tirole. 2009. Cognition and Incomplete Contracts. *The American Economic Review* 99, 1 (2009), 265–294.

[51] Oliver E. Williamson. 1975. *Markets and Hierarchies*. Free Press, New York.

[52] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2016. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. https://arxiv.org/abs/1610.08452

[53] Richard Zemel, Yu (Ledell) Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. *Proceedings of Machine Learning Research* 28 (2013), 325–333. Issue 3.