

Coercion: What is it and when is it justified?

James Slezak

Coercion is generally understood as the limitation of a person's freedom by some kind of threat. Pinning down a precise definition is more difficult however, and no-one claims to have done so. With this lack of a consistent formal understanding of coercion, laws and economic policies are nonetheless written to control coercive exchanges, largely based on intuitive judgments. An appropriate formal definition could place these policies on a firmer theoretical foundation, while properly capturing our intuitive understanding of the concept of coercion. This essay suggests a game theoretic approach to the problem that may avoid some of the difficulties presented by current approaches, while also revealing underlying connections to the concept of exploitation.

Since our intuition on the subject of coercion is in general relatively clear, a useful method of enquiry for evaluating claims about the nature of coercion is to test them against hypothetical scenarios. If a proposed definition is inconsistent with our intuitive judgment about whether an action is coercive or not, then we need to reject it or modify it somehow. Our aim is to find conditions that are both necessary and sufficient for coercion.

One of the clearest cases of coercion is the classic example of highway robbery. In this scenario, a robber offers his victim the choice "your money or your life," which we take to be a choice between being robbed and being assaulted. Unlike the victim of pick-pocketing who we would not regard as being coerced, the highway robbery victim in some sense chooses to hand over their money, but not in a free or just exchange. It seems reasonable initially to define coercion then simply as forcing someone to act by making a threat, but this definition immediately runs into problems, as noted by Basu (1990):

If Mr Buyer buys an orange from Mr Seller for a dollar, we could interpret this exchange as follows: "Mr Buyer threatens that if Mr Seller does not part with an orange, he will deny him a one dollar note." It follows that, given the above definition, the purchase of an orange (and, indeed, any good) would have to be described as an act of coercion.

If any choice can simply be rephrased as a threat, there must be some other feature that differentiates robbery from a free exchange. One possible distinction is that, unlike the orange buyer, both of the choices presented by the robber require the victim to relinquish some identifiable rights. Many cases of coercion involve the denial of some uncontroversial, easily defined rights, but other cases are not so clear. Taxation for example, just like robbery, is a means of separating people from their money using a threat. The threat of incarceration is a threat to curtail the right to freedom of movement, which would otherwise be regarded as a fundamental human right. So we are forced either to put taxation in the same category as robbery, or to invoke an increasingly complex notion of innate rights.

In fact, the denial of rights does not appear to be either necessary or sufficient for coercion. To show that it is not sufficient, we can imagine the case of a violent psychopath who offers his victim a choice between two different forms of assault. The psychopath offers a choice and denies rights, but whatever else we might say about the situation, we would be unlikely to call it coercion. In addition to being insufficient, we can see that the violation of rights is not necessary by considering the case of blackmail. Threatening to reveal some secret information about a victim does not necessarily deny them any rights, but we would still regard this as illegitimately coercive or exploitative behavior.

Finally, basing a definition on rights does not allow for the possibility of legitimate coercion. Threatening a home intruder with a weapon to force them to leave seems no less coercive than threatening a shop attendant with one. The relevant distinction is that we regard the former case as justified and the latter as unjustified. The fact that the tax office threatens to jail tax defaulters rather than having them assaulted does not make their conduct any less coercive, merely more justified¹. It appears that the normative question of the justifiability of particular threats is separate from, and must come after, the question of what actually defines a coercive act.

Nozick (1997), following Hart (1959), suggests using a complex set of criteria based on ideas originating in legal scholarship. Coercion is one of a very limited

¹ It could be argued that with prison conditions as they are in many countries such as the United States, the indirect threat of assault is actually already used.

set of circumstances that can invalidate legal contracts or exonerate someone from an otherwise punishable crime, so judicial pronouncements have at times touched on this issue. The American Law Institute offers the following definition of unjustified coercion, which they refer to as *duress*:

- (a) Any wrongful act of one person that compels a manifestation of apparent assent by another to a transaction without his volition, or
- (b) Any wrongful threat of one person by words or other conduct that induces another to enter into a transaction under the influence of such fear as precludes him from exercising free will and judgment, if the threat was intended or should reasonable have been expected to operate as an inducement.

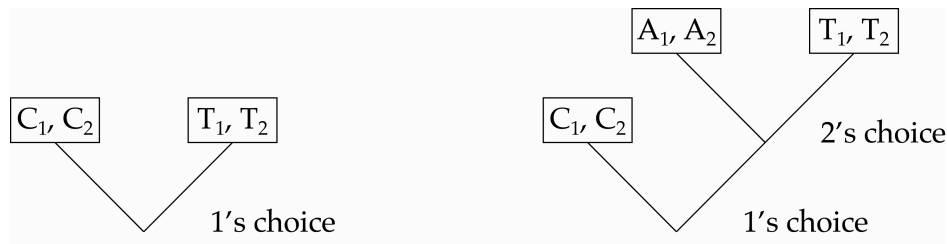
Both sections of this definition beg the question somewhat, with part (a) using the word *compel*, a synonym of *coerce*, and part (b) relying on a concept of “precluding [someone] from exercising free will”, which brings us no closer to understanding the conditions that might create such a situation. This leaves the judiciary relying more on an intuitive understanding of coercion than any explicit prescriptive definition, but this is presumably adequate for their purposes. Nozick proposes an alternative definition that avoids this shortcoming, but possibly introduces others. He suggests that a person *P* coerces a person *Q* not to carry out an act *A* if:

- (1) *P* threatens to bring about or have brought about some consequence if *Q* does *A* (and knows he’s threatening to do this).
- (2) *A* with this threatened consequence is rendered substantially less eligible as a course of conduct for *Q* than *A* was without the threatened consequence.
- (3) *P* makes this threat in order to get *Q* not to do *A*, intending that *Q* realize that he’s been threatened by *P*.
- (4) *Q* does not do *A*.
- (5) Part of *Q*’s reasons for not doing *A* is to avoid (or lessen the likelihood of) the consequence which *P* has threatened to bring about or have brought about.
- (6) *Q* knows that *P* has threatened to do the something mentioned in 1, if he, *Q*, does *A*.
- (7) *Q* believes that, and *P* believes that *Q* believes that, *P*’s threatened consequence could leave *Q* worse off, having done *A*, than if *Q* didn’t do *A* and *P* didn’t bring about the consequence.

Despite the careful wording and detailed argument leading to this definition, it appears to fail our original test of distinguishing between the orange-buyer and the highway robber. We can take P to be Basu's "Mr Buyer", Q to be "Mr Seller", A to be "keeping the orange" and P 's threatened consequence to be the withholding of one dollar. This definition would then categorize the buying of an orange as a coercive exchange. Nozick does suggest further refinements of his criteria that might resolve problems of this sort. However, without directly reviewing his other suggestions here, it seems unlikely that a concept on which fairly immediate intuitive judgments can be formed would require such a litany of subtle criteria in order to capture it formally. This is not to suggest that his approach is necessarily invalid, just that a neater and more direct definition would be preferable if one could be found.

An approach that offers the hope of achieving this goal is one that aims to characterize coercion in terms of game theoretic principles. If coercion could be explained as a particular type of strategy in a particular class of games, then the question of whether a given interaction constitutes coercion would reduce to questions about what game best models the real-life interaction at hand, and what strategy is being employed by one of the players. Such an approach might also hope to reduce the normative question of whether an instance of coercion is justified into easier judgments concerning the possible outcomes of the game.

In modeling an interaction where player 1 is coerced by player 2, the first observation to be made is that although it appears that only player 1 has a choice (as illustrated in the left-hand diagram below), in fact player 2 also has a choice, namely whether or not to carry out their threat in the case that player 1 does not cooperate. Player 2's coercive strategy then consists of ruling out, or appearing to rule out, the possibility of not carrying out their threat. In the diagram below, C stands for the coerced outcome, T for the threatened outcome, and A for the alternative option available to player 2 if they do not carry out their threat. In general, outcome A will be simply to walk away and have no further interaction.



Two models of a potentially coercive interaction

From this model, we can immediately deduce the inequality relations that must exist between the payoffs for each player. Clearly $C_2 > A_2$ and $C_2 > T_2$, otherwise player 2 would not be trying to coerce player 1 into choosing outcome C. For player 1, we must have $A_1 > C_1 > T_1$ for the threat to be effective and necessary. This leaves the relation between two of player 2's payoffs, A_2 and T_2 , undetermined. If $A_2 > T_2$, then player 2 can be said to be making a *vindictive* threat, since by choosing outcome T over A they would be acting against their own interests. In this case, player 2 needs to establish the credibility of their threat somehow, otherwise player 1 would not expect them to carry it out and would have no reason to cooperate with their demand to choose C. In the highway robbery example, we can assume that the risk of being caught makes player 2's payoff from carrying out the assault (T_2) lower than the payoff from walking away (A_2). Still, if player 1 is convinced that there is a significant chance that player 2 will carry out the threat T, then it is in their interest to cooperate. This analysis suggests why a successful robber will need to be perceived as irrationally prepared to punish non-cooperation, even at their own expense². It also explains how player 1 can try to avoid being coerced. If player 1 is known to have ruled out the possibility of outcome C in advance, then player 2 will choose A over T. Time-delay vaults in banks and signs in shops advising that "employees do not have access to the safe" are examples of this type of defensive strategy. On the other hand, if $T_2 > A_2$, which would be true in the example of a robbery taking place in a lawless land with a defenseless victim, player 2 can be said to be making a *credible* threat. In this case, no strategy can save player 1.

² This same strategy can often be seen in foreign policy.

In order for this game to be what we might call a *potentially coercive interaction*, there is one further condition on the payoffs. Namely C_1 (and also hence T_1) should be negative. That is, player 1 would suffer some loss by cooperating with the demand of player 2. The fundamental feature of this class of games is that one player, the victim, would rather not be playing in the first place. This suggests a simple definition of coercion: *any exchange such that if it were prohibited, at least one party would benefit*. The victim of robbery would clearly benefit if their exchange with the robber could not occur, but in the case of orange-selling, nobody would benefit from such a prohibition. This is in accord with the intuitive idea that in a free exchange we expect both parties to benefit. It can also potentially explain the other examples of coercion identified in the literature referenced below³.

In any game meeting the criteria of a potentially coercive interaction, we can say that player 2 employs a *coercive strategy* if and only if they rule out, or appear to rule out, an outcome A satisfying the inequality relations given above⁴.

Having proposed criteria for identifying coercion, we now look at the question of legitimacy. Deciding this question requires considering both the purpose of the threat being made as well as the proportionality of the outcome threatened, and inevitably requires judgments involving matters of degree. On the issue of purpose, we can see that apart from their methods, a relevant difference between the robber and the tax collector is where the money goes. If the robber were a corrupt police officer, he might be able to make the same threat as the tax collector (incarceration) and the same demand (handing over money), but would still not be acting legitimately since his purpose (theft) is illegitimate. On the other hand, the purposes of the government are relevant to deciding whether the coercion employed by its tax department is legitimate. It might impose taxation without representation, for example, or its officials might embezzle the money, or it might launch wars of aggression: there must be some circumstances under

³ In particular, it offers another, perhaps simpler, way of looking at the case of triadic interactions identified in Basu (1990).

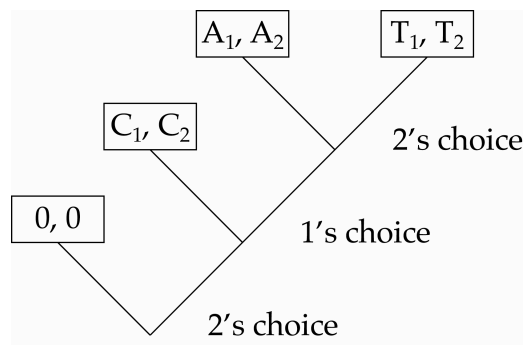
⁴ The word *strategy* in this sense involves both the game theoretic notion of a strategy (the choices that player 2 would make at each node) as well as the act of communicating this strategy to the other player — that is, threatening them.

which any given person would see tax collection as illegitimate. Somewhere between the robber and the tax department of a country with wholly agreeable policies lies a grey area in which the question of legitimacy is open to debate.

As for proportionality of the threat, the tax department in an otherwise well run government would also be acting illegitimately if it chose to enforce the tax code with capital punishment, since although its purpose would be legitimate (paying for necessary public services), its threat would clearly be disproportionate. Similarly, if you catch a co-worker stealing stationery from the supply cabinet, threatening to report them to the manager would be legitimate, whereas threatening them with grievous bodily harm would not be. This grey area between reporting a co-working and assaulting them, between jailing a tax defaulter and executing them, is the other dimension of the border between legitimate and illegitimate coercion.

In light of these observations, the best we can do is to reduce the question of legitimacy to smaller, potentially easier questions. If an interaction is modeled along the lines suggested earlier, we can say that a coercive strategy is legitimate if and only if we have the following conditions satisfied: outcome A is unjust, outcome C is just, and the threat T is proportionate⁵. We can define *proportionate* in this context more precisely as the condition that the payoff T_1 is as high as possible such that it is still less than C_1 . In other words, the only legitimate threat is one that is as lenient as possible while still being effective. If our means are unnecessarily harsh, they cannot be said to be justified by the ends. Determining whether A and C are just or unjust is more complicated, but to a first approximation, we can regard an outcome as unjust if one player gains at the other's expense — that is, if the payoff to player 1 is negative while the payoff to player 2 is positive.

⁵ A more general alternative would be to require only that C is *more* just than A , rather than using absolute terms.



A more complete model of a potentially coercive interaction

If we aim to prohibit or discourage coercion in some contexts, presumably the illegitimate ones, our model would be more complete if it formally incorporated the idea of “not playing in the first place”. This is the possibility that no exchange takes place, giving a payoff of zero for each player. “Prohibiting an exchange” then corresponds to mandating this null outcome, that is, not allowing player 2 to threaten player 1.

This more complete model suggests one possibility that has not been considered so far: an interaction in which the payoffs satisfy all the inequality relations specified for a potentially coercive interaction except that none is negative. To find such a situation, we could start with any coercive exchange and add some constant value to each of the payoffs. Since the strategies of the players depend on the relative utility of each option rather than any absolute value, we would expect them to make the same choices in this new game, for the same reasons. The question then is whether we still have a coercive exchange, in which case we would need to revise our definition, or whether the situation is now somehow qualitatively different.

As an illustration of this class of interaction, we might consider an employer in a developing country who offers a very low wage to a worker with no other available source of income. In the model proposed, outcome C corresponds to the worker accepting the low wage, and outcome T corresponds to the worker remaining unemployed. There will generally be an alternative option A available to the employer, which is to employ the worker at a higher wage that will still

allow the business to operate profitably. By ruling out option *A*, player 2 (the employer) forces player 1 (the worker) to accept outcome *C*. This strategy and the relations between the payoffs fit the definition of coercion given above except that player 1 would still be better off accepting outcome *C* than “not playing”. In fact, “not playing” is the same outcome as *T* in this case.

Although some people might use the word *coercion* to describe this situation, a more universally accepted term would be *exploitation*. Exploitation then can be seen as a generalization of coercion, and we can define it as any interaction that can be modeled by a game whose payoffs satisfy the inequality relations given earlier for coercion, without the requirement that any payoff be negative. This difference accounts for the intuitive qualitative difference between taking advantage of someone and robbing them.

As with coercion, there must be both legitimate and illegitimate forms of exploitation, and we would expect this distinction also to rely on matters of judgment and degree. Any profitable business can potentially afford to pay its employees more, but this does not mean that these are all cases of illegitimate exploitation⁶. If employment interactions are modeled by a game, then some notion of the underlying value of the work done (as opposed to its market value) must be introduced in order to specify the payoffs. Exploitation then could be regarded as illegitimate if the payoff (C_2) to player 2 is much greater than the payoff (C_1) to player 1.

Efforts to control illegitimate exploitation cannot be based on prohibiting interactions, since these interactions benefit both parties to some extent. Instead, if it is possible to ban the outcome *C*, we can expect the fairer outcome *A* to take place. Minimum wage laws are examples of this approach to the problem of exploitation. By prohibiting employment contracts involving very low wages, employers have no choice but to make fairer arrangements. Ideally, the benefits of an employment contract would be shared evenly between both parties in order to be completely non-exploitative, meaning that the rate of pay would to some extent reflect the profitability of the enterprise.

⁶ Marxists might argue otherwise.

The defining feature of both exploitative and coercive acts is the ruling out of an alternative that would otherwise be acceptable to both parties. This strategy is communicated to the other party, who is in the situation of having to make their move first. These interactions take place in situations where one party has the power to impose an undesirable outcome on another, and without some form of external regulation limiting the courses of action available, an unjust outcome can be expected. Understanding these interactions more thoroughly may provide more consistent and better means of formulating such regulation.

James Slezak

References

American Law Institute, 1981, *Restatement of the law second, contracts 2d*, St. Paul Minn.: American Law Institute Publishers

Basu, K., 1990, *Agrarian Structure and Economic Underdevelopment*, New York: Harwood Academic Publishers

Dutta, P., 1999, *Strategies and Games*, Boston: Massachusetts Institute of Technology

Hart, H., Honoré, A., 1959, *Causation in the Law*, Oxford: Clarendon Press

Nozick, R., 1997, Coercion, in *Socratic Puzzles*, Cambridge Mass.: Harvard University Press

Wertheimner, A., 1987, *Coercion*, Princeton: Princeton University