A Moral Theory of Free Will v2.0

Alexander Ash (a-ash@tuta.io)

July 2019, updated September 2020

This paper proposes an approach to normative ethics from the perspective of free will, emphasising the range and quality of volitions available to an agent as an indicator of liberty, morality as the absence of coercion, and empathy and apathy as virtue and vice, respectively.

Moral Agents & Patients

Agents are beings capable of intentionality (i.e., having a mental state for a desire, intention, or volition) and of carrying out actions.

Moral agents are beings with the ability to make moral judgments, which we'll consider here as any agent that surpasses a certain unspecified threshold of consciousness, rationality, and capability to distinguish right from wrong.

Moral patients are beings unable to make moral judgements, but that have a well-being and can be affected for better or worse by the actions of others.

Moral status (also called moral standing, or moral considerability) is the property of a being which characterizes it as having a right to moral consideration; a right to not be made to suffer, to not be killed, to have its well being taken into consideration, or to be treated in a certain way.

Science is unable to establish with absolute confidence the status of moral agency or patiency of a being, firstly because it is unable to compute with fidelity their quality of consciousness, rationality, and judgement ability, and secondly because the exact threshold of these variables which determine moral standing is currently uncertain and debated.

Therefore,[1]

1. Moral status is the degree of moral agency of a being.
2. Moral patients are a type of moral agent with low moral status.

The following is a non-exhaustive list of the moral status of various beings:

[1] We can also mention the vast differences in the levels of these three variables between mentally challenged humans, human members of MENSA, psychopaths, babies, chimpanzees, grasshoppers, parasites, and other beings, which would suggest different levels of moral agency between them.

- **Humans:** Rational humans have the highest degree of moral agency currently known, while non-rational or impaired humans (e.g., toddlers, the mentally disabled) are considered moral patients.
- **Non-human animals:** Under the 2012 “Cambridge Declaration on Consciousness” animals are acknowledged to have the capabilities for consciousness. Their ability to create and use tools, behavior in problem solving, and reasoning observed shows that there is non-rudimentary rationality in many species. Many animals can also differentiate between basic notions of right and wrong.[2] While these capabilities are not advanced enough to give animals moral agent status, it seems to be clear most hold significant moral status.
- **Plants:** While more controversial in status than non-human animals, there is evidence to suggest that plants have similar behavior to animals, can think and remember, choose between courses of action, gamble in response to risk, and are living beings which are aware of and respond to painful stimuli (although they seemingly can’t sense it as qualia). This evidence is challenged by some as merely indicating an automatic evolutionary response to stimuli. While their status as moral patients is disputed, we can make the argument their moral status is of enough significance to demand a basic consideration of their well being.
- **Information and communication technologies (in particular AIs):** The most advanced AIs in existence today could be said to be more akin to data processors than information processors, unlike humans. Paired with the ongoing debate regarding how to detect consciousness in AIs (and whether it is possible for them to develop consciousness at all), we can consider ICTs as possible future candidates for moral patiency and agency; currently however they are neither.

Free Will

A strong normative ethical theory is said to be one that all rational beings would endorse. Approaching this idea from a perspective of universality, we find that for there to be universal endorsing all rational agents would need to share the core components of the theory.[3]

[2] Most owners of pets would be happy to testify to this point, but there is no shortage of evidence. To give one example, a study describes dogs being happy to oblige to handshake offers by experimenters, but suddenly displaying distress signs when witnessing other dogs being rewarded with food after a handshake, implying they felt an injustice was happening beyond merely missing out on a treat.

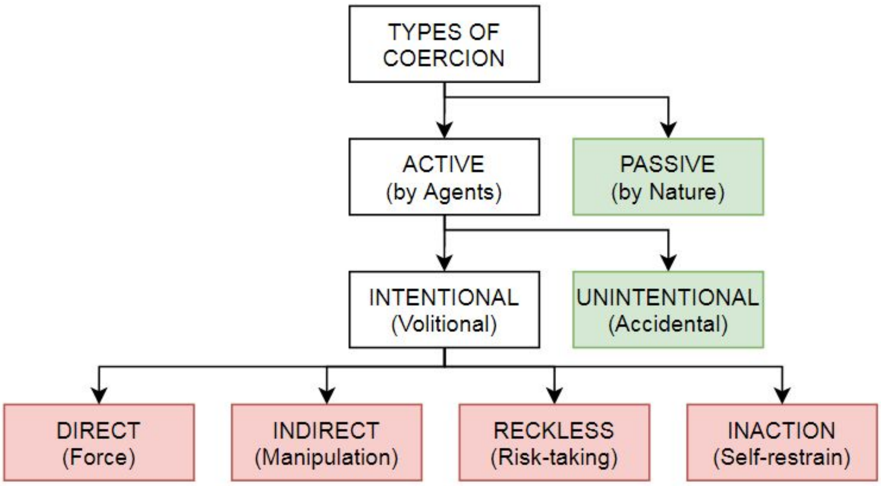
[3] Note that an agent can’t endorse that which he doesn’t understand.

This is important because we know that all groups have beliefs, values and practices that vary with their laws, religions, etiquette, and social contracts, while all individuals have a sex, race, and personality that is determined by their genetics and environmental upbringing. These are variables important to the response an agent has towards a moral dilemma, but they aren't fundamental to the dilemma itself. Moreover, intention (i.e., a reason for an action), property (i.e., possession of material), and duty seem to fall in this category as well,[4] and therefore will not be considered as critical to morality.

Consider however the idea of free will: an agent's range and quality of volitions open to him, with a volition being a unique decision taken consciously to effect or withhold an action. Because all moral dilemmas must involve a moral agent (by definition they are situations that require a moral choice, therefore a moral agent, and therefore volitions), the concept of free will shows itself as ever-present in morality.[5]

Coercion

We will define coercion as the restriction of an agent's volitions via some means; effectively the *constraint of an agent's free will*. Coercion is a spectrum, with different types of varying intensity. The following diagram shows this separation:



[4] Intention appears to be endemic to industrialized societies, while duty isn't universal by definition: It is an act or course of action that is required of one by position, social custom, law, or religion. Property is designed around a set of rules, and distributed via self-appointment, written account, or verbal account. While most societies will want to have some conception of property to prosperate, it would certainly not seem to fit as a universal.

[5] Volitions seem ideal exactly because 1) All moral agents (from humans to hypothetical extraterrestrials) have volitions, and 2) Volitions are something all moral agents can agree on wanting to carry out.

- **Passive:** Coercion imposed by nature (i.e., the material world and its phenomena), which prevents an agent from fulfilling its potential; e.g., a man drowning in a river, a woman with a failing kidney.
- **Active:** Coercion imposed by an agent, which acts as a restraint on other beings.
- **Intentional:** Active coercion resulting from a volition to coerce.
- **Unintentional:** Active coercion resulting from coincidence; e.g., a man accidentally dropping hot tea on a friend.
- **Direct:** Intentional coercion resulting from a volition which uses physical coercive actions; e.g., a man pushing, punching, or shooting another.
- **Indirect:** Intentional coercion resulting from a volition which uses non-physical coercive actions, such as threats, intimidation, or manipulation. Manipulation is the underhanded restriction of an agent’s volitions through deceptive or insidious means, like trickery or bypassal of reason. E.g., a woman guilt-tripping a man into buying her something. Indirect coercion often includes the threat of direct coercion (e.g., threatening torture).
- **Inaction:** Intentional coercion resulting from a volition to exert self-restrain and willingly allow harm; e.g., a man who chooses not to save a drowning child.
- **Reckless:** Intentional coercion resulting from a volition to disregard existing risk and not exercise restraint; e.g., a drunk person choosing to drive and running over someone. Note the original volition is *not* to coerce, but to effect an action that incidentally has a *chance* of collateral coercion. The risk of this collateral is taken, and thus it is intentional.

Negligence is not considered a universal type of coercion because it is based upon the societal construct of duty.[6]

The main separation of active vs passive coercion is tightly related to the idea of positive and negative liberty, where positive liberty is the possession of the power to fulfil one's own potential, and negative liberty freedom from external restraint by other agents. Therefore, *positive liberty is freedom from passive coercion, and negative liberty freedom from active coercion*.

[6] Negligence is a subset of unintentional coercion, governed by the concept of duty (which again is not universal by definition). Therefore, it is not an inherent type of coercion in nature, and any negligent action can be shown to be equivalent to the unintentional type (i.e., accidental) from an external perspective. For example, negligence in the workplace can entail failing to fulfill a duty of care based purely on social expectations, conventions, contracts, oaths, etc. Even situations that humans see as clearly negligent (e.g., a parent forgetting to feed his child) can be shown to be endemic - there are species of animals which abandon their babies and provide minimum or no care whatsoever. Gross negligence, defined as “a conscious, voluntary act or omission in reckless disregard”, is in this theory equivalent to reckless coercion, as there is a volition to disregard risk.

Therefore, if we understand coercion as the only element capable of decreasing free will, we can define *absolute morality as the complete absence of coercion*, both active and passive; in other words, as the complete opportunity to take control of one's life and realize one's fundamental purposes without the encountering of obstacles, barriers, or constraints.

To put all types of coercion into perspective, imagine a suicidal man on a bridge. An agent nearby could push him (direct), manipulate him into jumping (indirect), give him a pat on the shoulder disregarding the danger in doing so, making him fall (reckless), or placidly watch as he jumps (inaction). The agent could also accidentally trip and fall on the man, making him fall (unintentional), or a strong gust of wind could make the man slip and fall (passive).

Moral Responsibility

An agent is responsible for the coercion he inflicts. The amount of responsibility depends on the type and intensity of the coercion:

- **Passive (green):** Carries no responsibility. There is no responsible party, as nature is not an agent, nor can it willingly coerce.
- **Unintentional (green):** Carries no responsibility. There is no volition to coerce nor acceptance of risk (as in the reckless coercion); all coercion is a result of unforeseen accidents. The most common reason for unintentional coercion is human error, such as faulty memory, immaturity, lack of coordination, etc. The slightest belief that this coercion could occur changes the type of coercion to intentional (in particular to the reckless type).
- **Direct, indirect, reckless, inaction (red):** Carries responsibility. The more intense the coercion, the more the responsibility attributed. On average, some types of coercion are more coercive than others (direct > indirect > reckless > inaction.) More realistically, these types are spectrums with varying intensities, and can be equally coercive to one another. Direct coercion can be as minor as poking someone, but also as major as stabbing them to death. Indirect coercion can range from a minor insult to successfully inciting suicide. Reckless from dropping a hot beverage on a friend, to drunk driving and killing a pedestrian. Inaction from not warning someone they will trip, to letting someone drown in front of them.

In the example of the suicidal man on a bridge, most people would agree that pushing him (direct) is worse than manipulating him into jumping (indirect), which is worse than giving him a pat on the shoulder disregarding the danger in doing so, making him fall as a result (reckless), which is worse than watching as he jumps (inaction),

If an agent is coerced into coercing - such as those who are threatened, forced, pressured, manipulated, brainwashed, hypnotized, or forcibly drugged - then the responsibility for their acts is *proxied to the coercer*. For example, if Amy forces Bill to kill someone, then the responsibility of the death will fall on Amy, not on Bill. This is because Amy is enforcing her volition through Bill, while Bill is having his moral agency impaired while being coerced. In short, in order for there to be moral responsibility, an agent has to have the freedom to govern their desires (which goes beyond the mere freedom to act).

If an agent willingly coerces by request or order of another party (i.e., without them being coerced to do so), then the responsibility for the act falls on both parties. For example, if Alex tells Ben to kill someone, and Ben gladly obliges, then both Alex and Ben will have equal responsibility for the murder: Ben for the intentional coercion (murder), and Alex for proxying his volition to coerce through Bill.

Self-coercion is possible, with coercion again being the restriction of an agent's volitions, or the constraint of their free will (a person hitting themselves, shaming themselves, playing russian roulette, or doing nothing while their arm is on fire). Because the coercion is voluntary, there is no responsibility for the act.

Range of Responsibility

To what extent in time and space is an agent responsible for the coercion he exerts? Actions can have consequences continents away, decades down the line, many completely unexpected.

In short, his responsibility extends as far as his volition does, disregarding any unforeseen acts and coincidences that interfere with the causal chain.

Imagine I make a volition to playfully push my friend for having eaten my candy. I will be responsible for the push and possible scrapes he gets as a result (direct and reckless coercion respectively), but an unexpected broken leg due to the fall will be outside my range of responsibility, as my volition didn't include it (unintentional coercion). On the other extreme, a woman who saves a girl from dying is to be praised for her act, but not for incidentally saving another three children that she didn't know were also about to die.

As a consequence, the more aware of the effects of our actions we become, the more unintentional coercion will transform into reckless coercion (adding responsibility on top). A child running with scissors might unintentionally cut someone’s arm, sincerely having not considered the danger of their act. An adult running with scissors will almost always know that they are disregarding danger in doing so.

Thus, it is worth reiterating that even a slightest expectation of coercion when carrying out an act makes the type of coercion that occurs intentional. To put another example, If I offer a piece of cake to a friend knowing that it has a 99.99% chance of being delicious and a 0.01% chance of being poisonous (and delicious), it won’t matter if my intention is to make my friend enjoy some cake; my volition included the disregard of the risk of the cake being deadly (reckless coercion). Should my friend die, I will be responsible, although certainly to a lesser extent than if I had applied a more intense type of coercion.

Moral Encouragement

Because coercion is a spectrum, it follows that acts (a series of actions) also have degrees of encouragement or discouragement, rather than being imperative or prohibited.[7]

We defined morality as the absence of coercion (and as a corollary as the maximization of free will), therefore we can claim that the degree of moral encouragement present for an act is the *ratio of overall coercion prevented to coercion increased* due to said act. If the ratio is:

- Greater than 1, the act is *encouraged*. The greater the ratio, the greater the encouragement.
- Equal to 1, the act is neither encouraged nor discouraged. The amount of coercion prevented and increased is equivalent.
- Between 0 and 1, the act is *discouraged*. The smaller the ratio, the greater the discouragement.

[7] More arguments can be made in favor of responsibility as a spectrum. For example, the idea that a moral imperative would be embedded in nature and would therefore be impossible to disregard sounds unrealistic, and pushes us to toy with the idea of ranges. Another could be that if there was some absolute threshold between moral wrong and right, a single particle shifting position could change the imperative nature of an action in an instant. Moreover, a moral imperative would require us to act regardless of our own situation (it might require us to leave the hospital despite being sick, travel to dangerous places, face life or death situations, etc).

The ratio can tend to infinity when, for example, the coercion prevented asymptotes (e.g., all living beings are saved from perishishing) and/or the coercion increased tends to zero (e.g., all that is needed to save them is to will it). As the ratio increases, the encouragement for the act behaves evermore as an imperative. An inverse situation would make the ratio tend to zero, making discouragement behave evermore as a prohibition. Regardless, in both cases the choice of acting or not remains ultimately the agent's choice.

It is important to clarify that all actions included in the evaluated act should be causally tied. For example, if my act is "punch X, then save Y from drowning", we can imagine the ratio would yield a net positive result, yet punching X is very likely irrelevant to saving Y. Evaluating independent acts separately allows avoiding justifying the end via the means.

Note also that the coercion increased will never be exactly zero (making the quotient undefined), as there is always residual coercion included in any act, such as the restriction of our own free will by carrying out said act (i.e., the opportunity cost).

In practice it is not expected to use actual numbers to calculate ratios, but rather use the ratio as a quick heuristic tool to get a rough indication of the net morality of an act. This allows us to *describe* the encouragement or discouragement of a situation, rather than *prescribing* behavior. This theory denies the existence of true obligations or prohibitions, and claims we are perfectly allowed to dismiss high ratio courses of action and engage in low ratio courses should we so desire; however, we would be displaying a degree of "corruption" within our moral agency (for we would be making poor judgement calls) and a lack of virtue (discussed in a later section).

The ratio can be also used for the comparison of various courses of action. For example, helping a blind man cross the street is an encouraged act, but stopping a man from detonating a bomb in a nearby cafe would seem to be orders of magnitude more encouraged. In such a scenario, helping the blind man has practically no encouragement relative to stopping the detonation. Comparing to yet another act with a ratio even higher by many orders of magnitude - say, stopping a nuclear bomb from being detonated - would make the cafe bomb barely encouraged relative to this newer act. In short, encouragement and discouragement are relative to the available acts we have at our disposal.

Moral Justification

While coercion is an element to be minimized, some acts would appear to *require us* to coerce for the greater good. For example, most people would agree it is reasonable for a mother to push her daughter out of the way of an oncoming car, or for a policeman to take down a criminal who is about to shoot at a crowd.

The justifiability of an act is here tied to the encouragement (or ratio) of said act: For ratios greater than one, justification increases proportionally with encouragement. For ratios smaller than one, condemnation increases proportionally with discouragement. For ratios that are roughly one, acts are judged without any special consideration.

For example, killing a person to save a hundred lives would yield a very high ratio, encouraging the act and making the coercion exerted highly justified. Killing a man for no reason other than boredom would yield a very low ratio, discouraging the act and making the coercion highly unjustified. Killing a man who was about to commit murder would yield a ratio roughly close to one, adding neither special justification nor condemantion to the act. While the coercion applied is the same in all three scenarios, most people would lessen the guilt/punishment of the person who kills to save a hundred people, increase it for the person who kills for no reason, and allot “the usual” amount to the person who kills to prevent a killing.

Situation	Ratio	Encouragement	Justification
Kill 1 to save 100	> 1	High	High
Kill 1 for no reason	< 1	Low	Low
Kill 1 to prevent 1 death	~ 1	Neutral	Neutral

All acts carry some level of responsibility and justification. It should not be misunderstood that a high level of justification removes all responsibility from an act, or that a high level of responsibility removes all justification.

In situations where an agent is forced to make a decision between two or more coercive acts (e.g., trolley problem), the agent is not responsible for the choice as long as his volition is to minimize the coercion inflicted. This is because coercion is unavoidable one way or another, thus the agent is powerless in preventing it. If he however picks a course of action which leads to greater coercion than what was absolutely necessary, he will be responsible for the coercion inflicted (the difference between the coercion that was inevitable and the extra coercion exerted).

For example, if I can choose between saving one or ten humans, and decide to save the one, I will be responsible for the death of ten, but also for saving one, making me responsible for nine deaths (assuming all lives are of equal value). An argument can be made that the agent should be responsible for the full ten deaths, but in that case we would not be acknowledging that there was one life saved nonetheless.

Natural and Artificial Moral Dilemmas

Just like some acts appear to require us to coerce for the greater good, some would seem to require us to *refrain* from coercing for the greater good. For example, most people would agree it is unreasonable to kill a random person and harvest their organs in order to save two people in need of them at the hospital, despite the ratio of the act being positive (i.e., killing 1 random person to save 2). If a positive ratio was all that was needed to justify an act, then organ black markets would be morally justified.

We say an agent is *involved* in a moral dilemma if the reason for his participation in the dilemma is *not* due to pure chance or coincidence.

There are two types of moral dilemmas:

- **Natural dilemmas:** Those moral dilemmas where, excluding the decision-maker, all agents are involved. It is morally permissible to coerce for the greater good.
- **Artificial dilemma:** Those moral dilemmas where, excluding the decision-maker, at least one agent is non-involved. It is not morally permissible to coerce the non-involved for the greater good.

The artificial prefix is chosen because non-involved agents are dragged into a dilemma by the decision-maker, who considers whether to sacrifice them for the greater good -, despite the dragged agents having nothing to do with the situation.

Problem	Non-involved agent?	Natural/Artificial?	Answer?
Trolley Problem	No	Natural	Switch
Loop	No	Natural	Switch
Man in Yard	Yes, man in yard	Artificial	Don’t switch
Fat Man	Yes; fat man	Artificial	Don’t push the man
Surgeon	Yes, tourist	Artificial	Don’t kill the tourist

The result of the above rules is that in the Fat Man, Man in the Yard, and Surgeon scenarios, it is not morally permissible to kill the fat man, the man in the yard, or the tourist for the greater good, for they aren't involved. (If it was permissible, then everyone would live in fear of being killed for the greater good at any moment; think black market organ trading).

In contrast, switching tracks to the one with the lone man tied in the Trolley problem is permissible, as if the man is tied it means he is involved (regardless of whether he is aware of it or not, he was tied for a reason). In this theory, the Loop problem is equivalent to the Trolley problem; it makes no difference if the tracks reunite, the lone man is equally involved in both.

Awareness

As is often the case, the decision-maker must make a choice for another agent. For example, in the "Fat Man" dilemma he can choose to sacrifice the fat man, but it would be possible to let him choose himself whether to sacrifice his own life.

Understandably, there isn't always enough time to explain the situation to other people. If possible (i.e., as long as there's time and the situation allows it), choices involving the coercion of a third party for the greater good should be left to said party. Their choice should be respected, as they are effectively becoming the decision-maker.

If it is not possible to explain the situation/ask, the decision-maker must make an assumption: Either agents can be assumed to be selfish and care more about their own well-being, or selfless and care more about the great good. The choice of which assumption to make is up to the decision-maker, and both interpretations are understandable (service to self vs service to others).

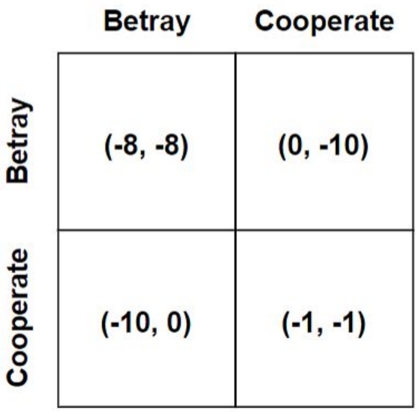
Virtue

Under this framework, being virtuous would equal having an absence of volitions to coerce. This makes for a strong argument that there might be one core virtue: empathy.

While definitions vary, here we refer to a specific type of empathy called *cognitive empathy*: The ability to accurately understand and identify with the mental state of other agents. This *does* not include affective empathy, the capacity to respond with appropriate emotions to another agent’s mental states (e.g., discomfort and anxiety in response to suffering).

As there cannot be understanding of an agent without knowledge of their mental state (both past and present), it follows that *perfect empathy demands perfect information*; an absolute understanding of people. Such a level of empathy would mean having a complete awareness of an agent’s circumstances, experiences, ideas, thoughts, beliefs, knowledge, desires, emotions, and volitions - and not only as a one-time snapshot, but continuously in time.

Empathy in our everyday life is said to make us more generous, kind, and unwilling to harm or use others, allowing us to reverse roles and experience reality from within the frame of reference of other agents. What would happen if this phenomenon was extrapolated to the extreme? My belief is that we would effectively identify with other people completely, extending our own self to encompass them. At such a point we would consider any fortune or misfortune to others as fortune or misfortune to ourselves in equal measure, and vice versa. Every action while interacting would take equal consideration of each side’s interests, which can be clearly reflected in the prisoner's dilemma, where two virtuous agents would naturally choose to cooperate with one another, thinking of the gain/loss of both parties involved instead of just one’s own. Each side would also have no doubt whatsoever at all times about their partner’s choice, as they would share a complete understanding of each other’s mental states.



	Betray	Cooperate
Betray	(-8, -8)	(0, -10)
Cooperate	(-10, 0)	(-1, -1)

Over time, such complete cooperation would overcome one of the biggest obstacles humans face when interacting with each other: trust. When mutual understanding takes place, cheating and lying no longer become possible *nor* necessary, barter and exchange are done fairly by default, inequality is tackled willingly when exposed to it, and shyness and social anxiety are no longer handicaps in expression.

Empathy could also bring out other virtuous elements in agents, like the courage to save a drowning girl, the kindness of charity for those in need, the honesty of character not to lie or manipulate, the respect of the natural rights of others, etc. This makes sense, as virtue tends to be tightly correlated with positive treatment towards other agents.

Moreover, if empathy means including others in our sphere of "self", any positive act towards others is a net benefit for ourselves as well. Cognitive empathy is compatible with the principle of the golden rule ("treat others as you wish others would treat you"), as by reliably understanding and identifying with others, the criticism that other agent's values or interests may differ from ours is ineffective.

Finally, an empathetic person will feel compelled to avoid coercing others (without justification, such as in positive ratio acts) due to the complete identification with their mental states. We can speculate a fully virtuous agent would therefore *never coerce intentionally without justification*, as there is an impossibility to disagree or have a volition to coerce.

But how can a virtuous agent encourage a change of opinion or volition in an agent without the use of coercion? Non-coercive alternatives indeed exist: *Influence* as an alternative to direct coercion, *persuasion* to indirect coercion, and *prudence* to reckless coercion.

We can define influence as the capacity to be a compelling force on an agent and affect his opinions or volitions (e.g., the influence of my father on myself when he interacts with homeless people), persuasion as the act of convincing to undertake or change the opinion or volition of an agent via rational argument or reasoning (e.g., a teacher persuading a student to follow her dreams and study art), and prudence as the capacity to exercise good judgement and common sense (e.g., being prudent and not driving home after drinking a whole bottle of vodka).

Unintentional coercion doesn't have a non-coercive alternative, and by definition can only be minimized by being more mindful of our actions and behavior; human error and chance are ultimately beyond our grasp.

Vice

If empathy is a virtue, is there such a thing as vice? Apathy fits as a good candidate. Apathy under this framework is the disregard and lack of accuracy in understanding the mental states of other agents, their circumstances, experiences, ideas, thoughts, beliefs, knowledge, desires, emotions, and volitions. In short, a perfectly apathetic agent is a fully selfish and self-centered person, without a care for the physical or mental state of others.

In contrast to the empath, an apathetic agent would utilize coercion wherever needed to reach his goals (one could imagine he would especially use manipulation, as it can pass unnoticed without creating trouble for himself). Similarly to empathy developing into other virtues, apathy would seem to develop into further vices: wrath, sloth, avarice - all of them based on the idea of the self (and usually having negative consequences for others), in contrast to virtues which are based on interaction with others.

Moral Responsibility vs Moral Judgement

There is a clear difference between moral responsibility and judgement that we should discuss.

If I try to hit someone and miss, or try to manipulate someone and fail, or drive intoxicated but get home safely, am I to be blamed if I actually never coerced anyone? After all, moral responsibility was related to the type of coercion inflicted, and in these cases no coercion was inflicted at all. Yet, we do see people in these situations often being blamed (e.g., attempted murder, driving intoxicated despite getting home safely, etc).

The answer is that while direct, indirect, reckless, and inaction coercion carry moral responsibility if they are successful, failed volitions to coerce carry only *bad moral judgement*.

Having faulty moral judgement is a sign of vice (i.e., apathy); a sign of immaturity, lack of care, and disregard for the safety of others. While this framework doesn't blame an agent for being apathetic, it insinuates his moral agency is deficient or in need of reinforcement. The agent might be in need of rehabilitation or moral education until his moral agency is restored.

Consider two drivers who skip a red light in recklessness. One of them is faced with a crossing child who he runs over despite trying to avoid him; the other is lucky and crosses safely. While the latter driver has faulty moral judgement, the former *also* has a moral responsibility for the (direct) coercion exerted.

In a more general sense, virtue and vice are indicators of the quality of moral judgements of an agent, while moral responsibility is concerned only with the actions taken by the agent in practice (i.e., acts and consequences).

Supererogation and Suberogation

While supererogation and suberogation are terms which typically require the introduction of duty (supererogation can be considered as “going beyond the call of duty”), we can make a brief commentary on them if we take care with our definitions:

1. Supererogatory acts are morally good, but are not required to be done (i.e., not wrong not to do).
2. Suberogatory acts are morally wrong, but are not condemnable to do (i.e., not right not to do).

A way to consider supererogation is to consider it as happening whenever coercion to oneself is voluntarily disregarded for the sake of other agents. More formally, when the ratio of an act is less than one, but by disregarding coercion to oneself it becomes greater than one. An example might be that of a homeless man who, despite risking starving to death, decides to give his last piece of bread to his beloved dog. Despite the act being discouraged due to the increase in coercion it generates to the man, it becomes a net gain if we accept such coercion for the sake of others. In other words, supererogation is self-sacrifice.

By this metric, charity is a supererogatory act only when there is a substantial loss happening, enough to push the ratio of the act below one. This means that a billionaire giving a few dollars to charity is encouraged, but it is the man who is in dire need of money who makes a supererogatory act by giving his last dollar away.

A suberogatory act on the other hand is more difficult to imagine, but it could be argued that it is one when an agent voluntarily decides to be coerced for his own benefit. More formally, when the ratio of an act is less than one, but by disregarding coercion to oneself for selfish reasons it becomes greater than one. An example of such an act might be that of a masochist who asks a sadist to beat him. Such an act is “not wrong not to do”, as the sadist enjoys it and the masochist disregards his own coercion.