強化学習によるメカトロニクス制御自動設計技術の獲得

Acquisition of Automatic Design Technology of Mechatronics Control through Reinforcement Learning

斎 藤 浩 一＊　菅 井 駿＊　桐 山 知 宏＊
Koichi SAITO　Shun SUGAI　Tomohiro KIRIYAMA

要旨

コニカミノルタは，「課題提起型デジタルカンパニー」としてお客様が持つ課題を自らが提起し，価値のある技術を通して解決し続けることで社会に貢献していく。

我々は，数十年後には高齢化社会の起因により，人口減少と働き手の高齢化によって深刻な働き手不足が引き起こされることを大きな社会課題として定義している。このような社会課題を解決するために，お客様に働きやすい環境を提供するメカトロニクス製品をタイムリーに提供し続けなければいけないと考える。これを実現するための要件は従来の制御ソフトウエアの設計手法では達成できず，新しい手法を用いた自動設計技術の獲得が必須であると考える。

自動設計技術を獲得するメカトロニクス製品の1stステップとして，MFP（Multi-Function Peripherals）本体の用紙搬送制御に強化学習手法（Reinforcement learning）を適用し，自動設計可能かを検証した。検証した結果，シミュレーション上で目的に応じた狙いの動作を強化学習によって設計することが可能であることを確認することができた。

本稿では，適用するまでに検討した内容を紹介する。

- 強化学習とはどういった手法なのか
- 強化学習を適用するために，用紙搬送制御では状態・報酬をどのように定義するか
- 用紙搬送制御へ強化学習を適用した結果

今後，実機を用いた検証を実施するとともに，MFP本体が持つ他の制御（画像印字制御・定着制御）にも適用範囲を拡大し，MFP本体制御全体に自動設計技術を適用していきたいと考えている。さらに，強化学習技術の適応範囲を広げ，働く人の目的に応じて最適に制御できるようなオフィス空間の構築を目指していきたい。

Abstract

As "A digital company with insight into implicit challenges", Konica Minolta contributes to society by continually presenting customers' problems and solving them through application of technology with value.

Several decades from now, due to an aging society, we will face a serious social problem that has been defined in terms of a shortage of working people due to a reduction in the population in an aging society. In order to solve this type of social problem, we believe it is necessary to continually provide in a timely manner mechatronics products providing customers with an easy environment to work in. Conventional control software design means are not capable of meeting the requirements to achieve this, so it is thought to be indispensable that we acquire automatic design technology that uses new methods.

As a first step of a mechatronics product on which an automatic design technology is acquired, we applied reinforcement learning to paper conveying control of MFP (Multi-Function Peripherals), and we checked whether automatic design would be possible. As a result, we confirmed with simulation that it was possible to design, by using reinforcement learning, an aimed operation corresponding to an object.

In this paper, we introduce what we studied before the application.

- What kind of method is reinforcement learning?
- How should we define conditions/rewards in paper conveying control in order to apply reinforcement learning?
- Results of applying reinforcement learning to paper conveying control

In the future, we will perform confirmation using an actual device and expand the range of control to which reinforcement learning is applied to other control functions of MFP (color printing control, fusing control), and would like to apply automatic design technology to all types of control throughout the MFP. In addition, we intend to expand the applicable range of reinforcement learning, and build an office space that can be controlled in the best manner, depending on the objects of the workers.

＊情報機器開発本部　エンジン制御開発センター　第2エンジン制御開発部

1　はじめに

働き手不足が今後数十年の期間で深刻化する社会課題に対し，課題提起型デジタルカンパニーの一員として課題解決に向けた取り組みを行っている。第2エンジン制御開発部のミッションとして，メカトロニクス製品における制御ソフトウェア開発を通してお客様に価値のある製品の提供を担っている。そのため，私たちの強みであるメカトロニクス製品を通じた，上記社会課題解決に取り組む施策を展開している。

近年，工場内のロボットや，建築現場での建機や，畑・田んぼでの農機を，機械学習を用いて工程作業に合った制御が可能なように制御ソフトウェアを構築し，作業者レスにて工程作業を実施していく技術の獲得が進んできている。当組織においても，今後開発していくメカトロニクス製品にてお客様へ価値のある動作をタイムリーに提供していくために，制御ソフトウェアの自動構築技術の獲得を加速して取り組んでいる。

施策の1stステップとしてMFP本体制御の用紙を搬送するメカ制御のソフトウェア開発に機械学習（強化学習）を用いることによって，用紙を搬送するための負荷（モーターやクラッチ）を駆動するタイミングを人手に依存せずに自動で設計可能か検討を行っている。本稿では，検討した内容・結果について紹介する。

2　MFP本体制御―用紙搬送制御

MFP本体制御では，お客様が印刷したい画像データ，原稿を，機内に設置している用紙に印字を行い，お客様の元へ印刷物を提供する。その過程の中で，主に3つの処理が動作し機器を制御する。

1つ目は，所定の位置から所定の位置まで用紙を搬送する用紙搬送制御。2つ目は用紙が所定の位置まで到達したら，用紙に画像を印字する画像印字制御。3つ目は，用紙に印字した画像を定着させる定着制御である。今回の取り組みは，用紙搬送制御をピックアップし，適用していく。

2. 1　用紙搬送制御の目的

当社製品の用紙搬送制御の目的は，用紙を設置しているカセット（トレイ）から用紙を給紙し，機内に設置している複数のローラーにて用紙を挟持（ニップ）し搬送し，搬送過程の中で用紙にダメージを与えることなく機外へ排出することを目的としている。ここで，用紙にダメージを与えてしまう例としては，所定のタイミングにてローラーを駆動できなかった場合，停止中のローラーに用紙が衝突してしまい用紙が折れてしまう，また搬送中の用紙同士が衝突してしまうと同様に用紙が折れてしまう，さらに用紙を複数のローラーでニップしている場合，ローラー間の速度差によって不要な用紙の撓み状態を作ってしまうことである。また，用紙にダメージなく搬送することに加え，所定時間の中で所定の枚数を排出可能にすることを目的としている。

3　強化学習（Q-Learning）

3. 1　概要[1)]

強化学習とは，試行錯誤を通じて制御対象に適応する制御方法を学習する手法である。制御対象がある状態時に任意の動作をさせた結果，変化した制御対象の状態より，制御対象が達成すべく目的に向かうことができたかに応じて，動作の良し悪しを判断し報酬を与える。与えられた報酬値に応じて，制御対象がある状態時に行う動作の価値（以降，行動価値）を学習する。この過程を繰り返し学習していくことで，最終ゴールを達成可能な制御方法を習得していく。Fig. 1 に強化学習の学習概略図を示す。

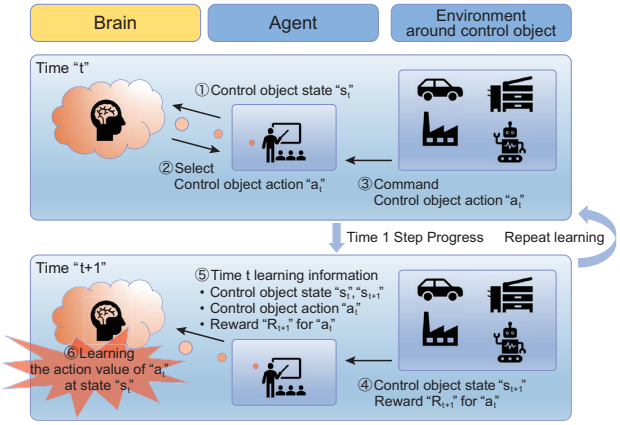

Fig. 1 The framework of reinforcement learning.
Reinforcement learning is a method of learning, through trial and error, an appropriate control method for controlling a control object. Learning is controlled by three blocks (brain, agent, and environment). Action value is calculated from a reward obtained as a result of action of an arbitrary action. By repeatedly carrying out this process, a control method capable of achieving the final goal is learned.

学習は3つのブロックによって制御される。1つ目は，学習を制御する学習者（Agent），2つ目は，学習した結果，制御したい対象を取り巻く環境（Environment），3つ目は，学習していく過程を学習・記録していく頭脳（Brain）である。Fig. 1 記載の強化学習の学習シーケンスを説明する。

任意の時刻t（$t \geqq 0$）時に，

①学習者（Agent）は，時刻tの制御対象（Environment）の状態s_tを取得し，頭脳（Brain）に伝達する。

②頭脳は，制御対象状態s_t時に選択すべく動作a_tを任意の選択方法（探査戦略）に基づいて決定し，学習者へ伝達する。

③学習者は，取得した制御対象の動作a_tにて動作するように制御対象へ指示する。指示を受けた制御対象はa_tに基づき動作する。

学習周期1ステップ後（時刻t+1）に，

④学習者は，時刻t+1の制御対象の状態s_{t+1}と，制御対象の状態s_t時に動作a_tにて動作した結果，制御対象を目的に応じた状態に遷移することができたかの判断結果に基づき報酬R_{t+1}を取得する。

⑤学習者は，時刻t時の学習に必要な情報：時刻t時の制御対象の状態s_t，制御対象の動作a_t，時刻t+1時の制御対象の状態s_{t+1}，状態s_t時に動作a_tを行った際に取得した報酬値R_{t+1}，を頭脳へ伝達する。

⑥頭脳は，取得した学習情報を用いて，時刻t時の制御対象の状態s_t時に，動作a_tを行う行動価値を演算し，記憶する。

⑦①～⑥を各時刻で繰り返し試行錯誤し，制御対象の状態に応じた最適な動作を行動価値に基づいて決定していく。

上記学習シーケンス内にて行動価値を算出に辺り，以下の式Fig. 2を用いて算出する。

$$Q(s_t,a_t) = Q(s_t,a_t) + \eta * \left(R_{t+1} + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t) \right)$$

η : Learning coefficient
γ : Time discount rate

Fig. 2 The formula for action value calculation.
$Q(s_t, a_t)$ is the action value when an action 'a_t' is taken in a state "s_t" at time "t". R_{t+1} is the reward for the action. The action value has the following characteristics. The higher the reward for the action is, the higher the action value is evaluated. In addition, when it is highly valuable to reach the next state, the action value is evaluated to be high, and when it is valueless to reach the next state, the action value is evaluated to be low.

Fig. 2は，時刻t時の状態s_t時に動作a_tを取る行動価値をQ（s_t，a_t）で表している。また，行動価値を算出する過程における特徴としては，第2項目で表している次の状態の行動価値の最大値と現在の状態での行動価値の差を取っている箇所にある。これは，次の状態に到達することに大きな価値がある場合，時刻tに取った状態s_tと動作a_tに対する行動価値が大きく評価される（現状の行動価値よりも大きく算出される）ことにある。次の状態に到達することに価値がない（次の状態の行動価値が低い）場合，s_tとa_tに対する行動価値は低く評価される（現状の行動価値より低く算出される）。この数式を用いることにより，最終的に行動価値が高い状態sと動作aを導き出すことが可能となり，制御対象が目指すべき目的にあった制御を導き出すことが可能となる。

3. 2　強化学習のメリット

このように学習を行う強化学習を適用することのメリットは2点あると考える。

1点目は，制御対象の不確実な状態変化を考慮に入れることなく，制御規則を習得することが可能なことにあると考える。ここで不確実な状態変化とは，制御対象内のユニット（搬送経路，モーター，ローラー，etc）の性能変化のことを指している。強化学習は，制御対象が試行錯誤して動作した結果，目的を満たした状態になることができたか，また目的を満たさない状態になってしまったかを判断し，報酬値として状態に応じた動作を評価することで，最終ゴールを目指す学習手法である。そのため，設計者は，制御対象をどのような目的で動作させるか，どのように目的を達成しているかの判断を考慮に入れるだけで，不確実な状態変化を含んだ制御規則を構築することが可能である。

2点目は，設計者の作業量においても少なくて済むことにあると考える。設計者は様々な制御対象の動作タイミングや制御状態の組み合わせを考慮することなく，目的と目的達成を判断する方法を構築するだけで，学習者が試行錯誤して自動で目的にあった制御を構築していくことが可能なため，従来手法（タイミングを考慮した制御仕様の構築）よりも設計すべき作業量が少なくなる。

4　用紙搬送制御への強化学習適用

強化学習をMFP本体制御へ適用することにより，自動でお客様の求める製品動作を提供する制御技術の獲得に取り組む。ここで，MFP本体制御内において，複数ユニットの連結で構成される用紙搬送制御ではユニットの性能変化により目的を達成できなくなってしまう可能性が高い。例としては，1つの機体の中で複数のローラーが存在し，ある1つのローラーが摩耗状態に応じて単位時間当たりの用紙の搬送量が低下してしまい，設計値とは異なる搬送量になってしまう。これが，発生してしまうと，用紙ジャムや排出できたとしてもダメージが発生した状態で排出してしまう。そのため，強化学習のメリットを最も活かせる制御として用紙搬送制御への適用を検討した。検討にあたり，以下の4点の項目の検討を行った。

4. 1　シミュレーター採用の検討

強化学習の特徴として，どのような動作を行ったらよいか分からない状態で，試行錯誤を繰り返して制御対象の状態に応じた最適な動作を決定し制御を構築していくため，実機を用いて学習を行うと実機を破損してしまう可能性がある。そのため，今回の検討では実機同等の振る舞いをするシミュレーターを用いて検証を行った。ここで，実機同等の振る舞いとは，速度を指定すればモーターが狙いの速度で駆動する，またモーターの駆動力をローラーに伝えるか伝えないかを切り替え可能（連結or非連結）なクラッチをモーターと接続し，ローラーの駆動を切り替え可能な振る舞いが可能であることを示す。今回の検討では，強化学習が用紙搬送制御に適用可能なのかを即座に確認したいため，制御対象として製品構成を簡易化した構成を用いて検証を行った。Fig. 3に簡易化した構成を示す。

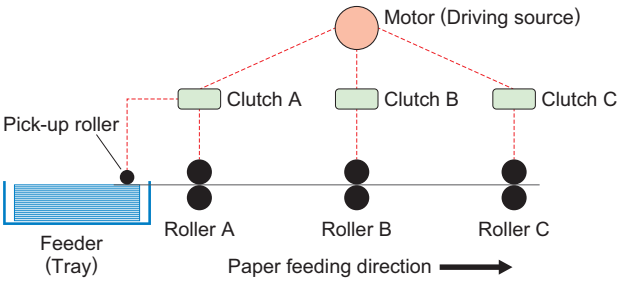

Fig. 3 The configuration of the control object, using simulation.
The motor and the clutches are controlled at arbitrary timing to drive the rollers so that paper can be conveyed from the paper supply tray and be ejected outside the machine.

Fig. 3 の動作概要を説明する。モーター（Motor）が任意の速度で駆動している場合，各クラッチ（Clutch A, B, C）が連結状態である場合，各ローラー（Roller A, B, C）はモーターの速度と同じ速度で駆動する。給紙トレイ（Feeder）はローラーAとクラッチAを介してモーターの駆動力を共有し，モーターが駆動状態時にクラッチAを連結すると用紙の給紙を開始する。給紙した用紙は，ニップしているローラーが駆動していれば，単位時間あたりの速度で搬送される。モーター，各クラッチを任意のタイミングにて制御することによりローラーが駆動し，給紙トレイから用紙が搬送され機外へと用紙を排出することが可能となる。

4. 2　目的から強化学習に学ばせる項目を検討

用紙搬送制御の目的から強化学習に学ばせたい項目を定義する。用紙搬送制御の目的から，ポイントとなる観点を①用紙にダメージを与えることなく，②所定時間内で所定の枚数を機外へ排出することを強化学習で学ばせたい項目として抽出した。①・②のままでは，抽象化された観点であるため，用紙を搬送する上でどのような状態になることが目的を達成しているのか具体化して定義した。以下に具体的な状態を示す。

①用紙にダメージを与えることがない。
- I.　経路内で搬送中の用紙同士が衝突しないこと
- II.　経路内で搬送中の用紙傾きを補正できること

②所定時間内で所定の枚数を機外へ排出する。
- III.　用紙を機外へ排出できること
- IV.　所定時間内に狙いの枚数排出できること

上記4つの観点を満たす制御を構築することで，目的を満たすことが可能と判断した。

4. 3　制御対象の状態と動作の検討

現状制御設計を元に，制御対象の状態s，所定の状態s時に取る動作aをどのような情報にするかの検討を行った。用紙搬送制御の現状制御は，製品仕様値を達成すべくFig. 4 のような設計を行って，各負荷を動作させるタイミングを設計する。

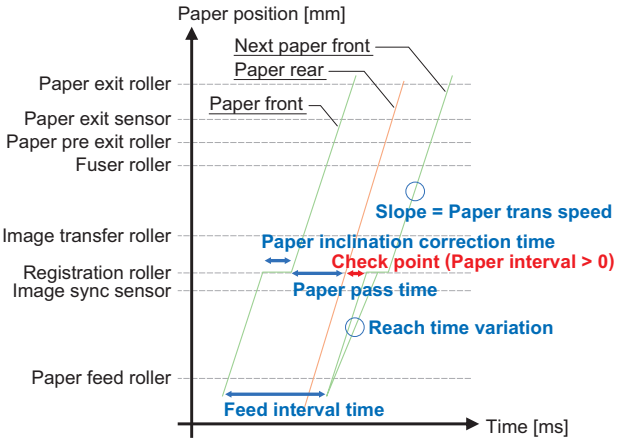

Fig. 4 A paper conveyance timing diagram in the case of conveying A4 size paper.
The diagram illustrates when and where the top end and the tail end of A4 size paper should be located to convey the paper in a targeted manner.

Fig. 4 は例としてA4サイズの用紙を搬送した際に，用紙の先端と後端がどのタイミングで，どの位置にあれば狙いの用紙搬送が可能かを表した用紙搬送ダイアグラムである。例として，位置決めローラー（Registration Roller）にて，用紙後端と次用紙先端の間隔（紙間）が確保できている（0より大きい）ことを達成するために，用紙の搬送速度，位置決めローラーの通過時間，次用紙の給紙時間，給紙口から位置決めローラーの到達ばらつきを考慮に入れて，“用紙後端位置”・“次用紙先端位置”がどのような位置にある際に，用紙搬送するための負荷を動作する必要があるか設計する。ここで，ポイントとなるのが，用紙位置（先端／後端）に応じて負荷を動作させるタイミングを決定することである。そのため，本検討において制御対象の状態を“用紙位置（先端／後端）”，動作を“各負荷の動作パターン”とすることで学習が可能か検証を行った。

また，用紙搬送制御では0.5mm単位の精度を必要とする。精度を荒くしてしまうと，目的を満たす制御を実現することができない。そのため，本検討においても，0.5mmの搬送精度を満たすために，学習1ステップ単位を用紙の搬送速度に応じて設計し，学習周期にて制御対象の状態を監視し，状態に応じた動作を切り替えて検証を行った。

- 学習周期
  1 ms
  ※搬送精度：0.5mm，搬送速度：166.4mm/s
  0.5mm÷166.4mm/s→3.004ms>1msとなり，搬送精度を満たすことが可能。
- 制御対象の状態s
  経路内に存在する用紙位置（先端／後端）
- 制御対象の動作a
  ローラーA，ローラーB，ローラーCの動作パターン（Table 1）

Table 1 The relationship between the action pattern and the rollers' states.

Action pattern	Roller A	Roller B	Roller C
Type 1	Rotation	Rotation	Rotation
Type 2	Rotation	Rotation	Stop
Type 3	Rotation	Stop	Rotation
Type 4	Rotation	Stop	Stop
Type 5	Stop	Rotation	Rotation
Type 6	Stop	Rotation	Stop
Type 7	Stop	Stop	Rotation
Type 8	Stop	Stop	Stop

4. 4　報酬の検討

各用紙位置に対するローラーの駆動パターンの行動価値を学習するために，用紙搬送制御の目的（I. ～ IV.）を満たししているか判断し，結果に応じて報酬を与える。本稿では，各目的に応じて報酬値を与える判断と与える報酬値について説明する。また，報酬の与え方によって発生した課題と対応した内容を説明する。

4. 4. 1　各目的に応じた報酬値と与える判断

各目的に応じて与える報酬値として，目的を達成できた状態と行動に対しては“正（R>0）”の値を持つ報酬値を与え，行動価値を高めて良かった行動として記憶する。また，目的を達成できなかった状態と行動に対しては“負（R<0）”の値を持つ報酬値を与え，行動価値を低くして悪かった行動として記憶する。このような値で報酬を与えることによって，良かった行動，悪かった行動を学習し，ある状態において，目的を満たした行動を選択することを可能とした。

次に，目的に応じてどのように報酬を与えるかの判断を説明する。

I.　経路内で搬送中の用紙同士が衝突しないこと
学習周期単位で経路内に存在する用紙の先端位置と後端位置を監視し，前用紙の後端位置よりも次用紙の先端位置が先行してしまった（前用紙と次用紙が衝突を起こした）場合，悪かった行動として報酬値“−1”を与える。

II.　経路内で搬送中の用紙傾きを補正できること
用紙が所定位置に到達した際（用紙傾きを補正する位置）に，学習周期単位で用紙傾きを補正する動作が正常に行われていなかった場合に，悪かった行動として報酬値“−1”を与える。ここでFig. 5に用紙傾きを補正する動作について示す。

Fig. 5に示す用紙傾き補正制御は，ローラーBに到達する際に制御用紙の傾きを補正する動作を制御する。用紙先端がローラーBに到達する前に，ローラーBの駆動を停止，ローラーAを駆動することにより，用紙先端をローラーBに押し当てることで用紙傾きを補正する。また，押し当てる量（用紙を撓ませる量）に応じて，ローラーAを駆動する時間を制御する。今回の検証では，上記タイミングにてローラーBが停止しているか，ローラーAが狙いの時間分駆動できているかによって判断を行った。

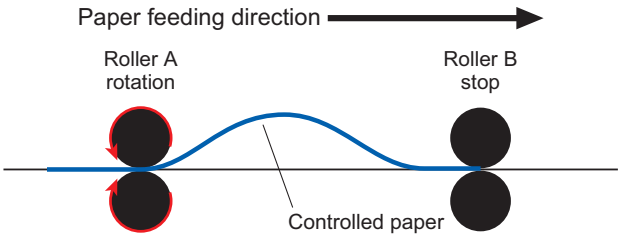

Fig. 5 Paper skew correction control.
Paper skew correction control controls an action to correct the paper skew when it reaches the roller B.

III.　用紙を機外へ排出できること
学習周期単位で経路内に存在する用紙の後端位置を監視し，後端位置がローラーC位置を超えた場合，良かった行動（目的を達成できる行動）として報酬値“+1”を与える。

IV.　所定時間内に狙いの枚数排出できること
III.と同タイミングにて1枚当たりの排出時間が狙いの時間以内だった場合，良かった行動（目的を達成できる行動）として報酬値“+2”を与える。

4. 4. 2　発生した課題と対応

IV.の学習を行っていく過程で，目的である所定時間内に狙いの枚数排出することを達成することができないという課題が発生した。当初は，IV.の報酬を与える判断を異なった内容で行っていた。報酬を与える判断として，所定時間経過後に排出できた枚数が狙いの枚数以上だった場合，報酬値“+2”を良かった行動として与えていた。しかし，目的を達成することができなかったため行動価値の遷移を分析した。

行動価値の遷移分析の結果，良かった行動として報酬値を与えられた後に報酬をもらわない状態を継続すると，報酬値によって高まった行動価値が低下してしまい良かった行動を記憶しておくことができないことだと分かった。Fig. 6に行動価値の遷移を示す。

同図は縦軸を用紙がローラーCの1つ手前に存在する状態時にローラーCを通過可能な行動における行動価値の遷移を示す。また，横軸を試行1回内の学習ステップ数を表している。学習ステップ61のタイミングで報酬値“+1”を与えることにより，行動価値が学習ステップ60よりも高くなっているが，報酬を与えない状態を継続することにより，学習ステップ65では行動価値が報酬を与える前の値と同等まで低下していることが分かる。そのため，行動価値の低下の影響を軽減するために，報酬を与える判断タイミングを増加し，良かった行動を記憶し続けるように報酬を与えるタイミングを増加させる必要があると判断した。

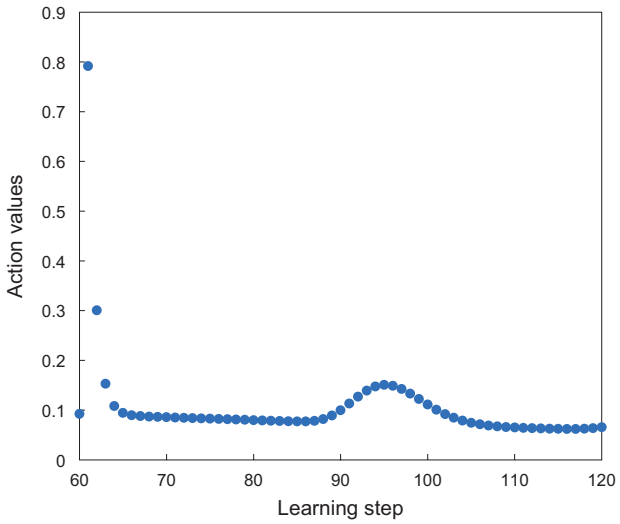

Fig. 6 The transition of the action value.
The vertical axis represents the action value in an action that enables the paper to pass through the roller C in the state where the paper is one step before the roller C. The horizontal axis represents the number of steps in one trial. When a state where no reward was given continued for a while after a reward value was given for a good action in step 61, the action value once raised by a given reward decreased, and the good action was not memorized. From such a result, we determined that it was necessary to increase the number of timings at which a reward was given so that a good action was kept memorized.

5 適用結果

検討内容を考慮に入れて用紙搬送制御へ強化学習を適用した結果を示す。Fig. 7 に結果を示す。

同図は，縦軸が学習過程において所定時間内の用紙排出枚数（PPM：1分間当たりの排出枚数），横軸は学習試行回数（episode）を示す。学習単位として，「試行1回あたりの時間を1分間として，10,000回の試行を繰り返す」を1回のカリキュラムとして学習を行った。①は，未学習状態から1回のカリキュラムを学習した結果を示す。②は，①の学習した結果を引き継いで，6回カリキュラムを学習した結果，③は10回学習した結果を示す。④は，①・②・③を1つにまとめたグラフを示す。①の結果から，学習初期はPPM=0であり用紙を1枚も排出することができていない。しかし，I. ～ IV. の目的に沿って，試行錯誤しながら学習を複数回行うことで，報酬が与えられて用紙を排出することが可能となってくることが分かる。また，①～③の結果より，用紙排出枚数を増加させるために動作方策を試行錯誤しながら切り替えるため，用紙衝突／傾き補正ミスが発生し，学習していた排出枚数よりも少なることが分かる。この過程を繰り返し行うことで，④の結果からわかるように学習を継続して行うことで，用紙排出枚数が増えて最終ゴール（目的のPPM）に近づいてくることが分かる。また，学習結果から構築された制御を用いて用紙搬送制御をシミュレーション環境にて実施した結果，学習した通りに用紙搬送制御の目的に沿って用紙を排出することが可能となった。

これらの結果より，特定の環境下（今回使用したシミュレーター環境）では用紙搬送制御に強化学習を適用し，人手に依存せずに用紙を搬送するための負荷を駆動するタイミングを自動設計可能であると判断する。

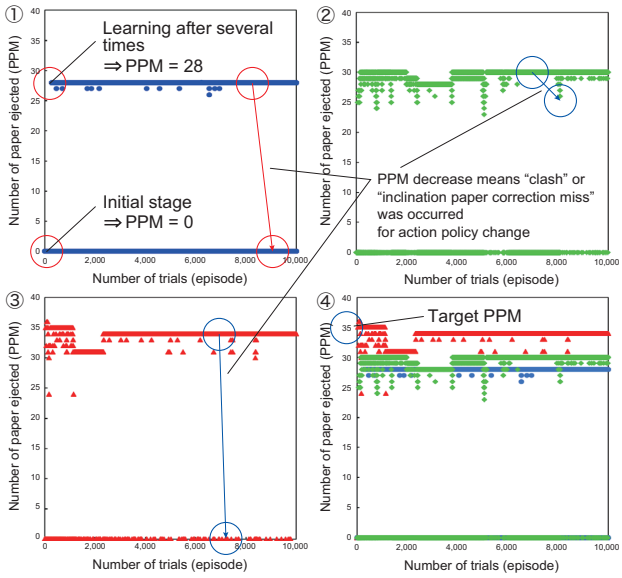

Fig. 7 The transition of the number of ejected paper sheets at the time of learning.
①: First curriculum
②: Sixth curriculum
③: Tenth curriculum
④: Diagram of the superposition of ①, ②, and ③
The horizontal axis represents the number of learning trials (episodes), and the vertical axis represents the number of ejected paper sheets per minute (PPM). One episode takes one minute, and 10,000 episodes repeated in succession constitute one curriculum. By continuously carrying out the learning, the number of ejected paper sheets increased and became close to the target number (35 PPM).

6 今後の展望

今回使用したシミュレーター環境は，実機が持つ不確実性のある状態を表現できていない。そのため，用紙搬送制御が考慮に入れなければいけない不確実性の状態を学習可能であるか判断できていない。今後，実機を用いた検証を実施し，強化学習のメリットを最大限活かすことができるか確認していく。また，MFP本体が持つ他の制御（画像印字制御・定着制御）にも適用範囲を拡大し，MFP本体制御全体に自動設計技術を適用していきたいと考えている。

これらの技術検証と並行して，自分たちが目指す世界の構築にも取り組んでいきたいと考えている。今後数十年の期間で，高齢化社会が引き起こす働き手不足が深刻化すると考えている。そのため，オフィス（すべての働く空間）での働き方も大きく変革する必要があると考える。こういった社会に対し，どのような年齢層でも意識することなく，個人が思い描いたように働くことが可能な社会（個が輝く世界）を目指していきたい（Fig. 8）。

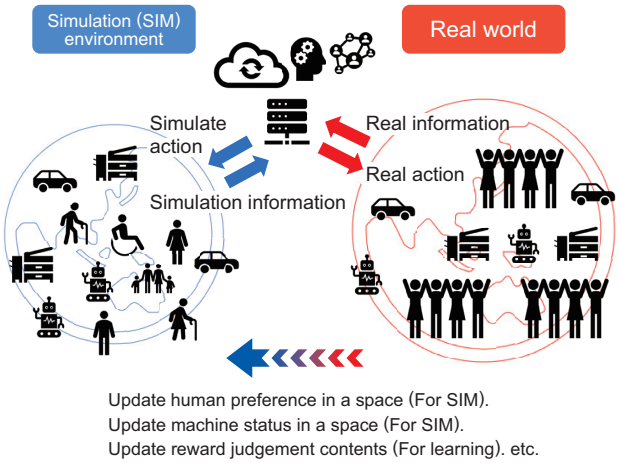

Fig. 8 A space where mechatronics products can be most appropriately controlled depending on the objects of working people.
Taking the office environment as the control object, in order to reach the goal of providing an environment where workers and mechatronics products can comfortably work, the way how the workers and the mechatronics products in the space are controlled is automatically configured by performing trial and error and giving rewards depending on objects.

そのような社会を実現するために，オフィス環境で価値のある動作を提供可能なメカトロニクス製品をタイムリーに提供することが必要である。また，これらのメカトロニクス製品が働く人の目的に応じて最適に制御できるような空間を構築することが必要であると考えている。

このような空間制御を構築するために強化学習が適用できるのではないかと考える。オフィス空間を制御対象と捉え，働く人とメカトロニクス製品が，働きやすい環境の提供という目的を達成するために，どのように空間内の働く人・メカトロニクス製品を制御すればよいか，様々な環境で試行錯誤して学習し，目的に応じて報酬を与えることによって自動で制御を構築していく。これを達成することによって，個が輝く世界を達成し，今後の社会課題を解決していきたいと考えている。

●参考文献

1) 小川雄太郎（2018）『つくりながら学ぶ！深層強化学習』マイナビ出版